INFT 4836: Introduction to Big Data Analytics

Basic Information

- **ADA University, Fall Semester, 2017**
- **INFT 4836: Introduction to Big Data Analytics (6 credits)**
- **Course meeting times and location:**
  - Section A: Monday/Wednesday, 10:00-11:15 B-301
- **Instructor:** Abzatdin Adamov (PhD in Computer Science)
- **How to contact instructor**
  - In-person office hours: **drop-in anytime between 9:00-12:00 am on Tuesday or Thursday in SPIA 316B.**
  - E-mail addresses: aadamov@ada.edu.az
  - Phone numbers: **(012) 437-32-35/ 335**
  - Preferred mode of communication: **E-mail**
  - Optional: **drop me an email anytime so we agree on a meeting**
- **Course Web page URL:**
  Section A: TBD

Course Description

- **Prerequisites:** Prerequisites will not be applied for this course, but I suppose that you successfully passed or have knowledge/experience in following courses: CSCI 2406 - Computer Organization & Architecture, CSC 105 - Programming Principles I, CSCI 3615 - Database Systems, CSCI 2303 - Introduction to Computer Networks

- **Technology requirements:**
  - Equipment: Students are encouraged to use their laptops to install required software, do appropriate platform settings, implement the class assignments and implementations.

- Software: R and RStudio will be mostly used for Data Manipulation and Analytics. Limited time will be allocated for learning it why having programming skills are important (for some examples we'll use Java also).

- **Overview of course (as stated in the program catalogue):**
  The Internet Services, Web and Mobile Applications, Pervasive Communication widely available today that are meeting many of our needs have stimulated production of tremendous amounts of data (call metadata, texts, emails, social media updates, photos, videos, location, etc.). Even with the power of today's modern computers it still big challenge for business and government organizations to manage, search, analyze, and visualize this vast amount of data as information. Over 90% of this information is unstructured, what means data does not have predefined structure and model. Generally, unstructured data is useless unless applying data mining, data extraction and advanced data analytics techniques. At the same time, just in case if you can process and understand your data, this data worth anything, otherwise it becomes useless.

  Big Data Analytics is scientific process of transforming data into knowledge large amounts of data into actionable knowledge (intelligent insights) enabling Data-driven decision making. Big Data Analytics is not brand new technology and existing many years before. But because of numbers of factors all come together now, this technology is becoming more and more important today.

  In "Introduction to Big Data Analytics" is designed to provide students with fundamental knowledge on: reasons of Big Data problem, use-cases of Big Data by multi-sectoral industries, distributed architectures and platforms for Big Data storing and processing. Students will be introduced with architecture of Hadoop, HDFS and concept of MapRecuce. Several other key components of Hadoop Ecosystem will be introduced as well. Real data analytics and visualization will be accomplished using R programming language.
- **Student learning objectives (as stated in the program catalogue):**
  1. Understanding Big Data nature and what drives Big Data
  2. How Big Data Analytics can effect business and bring them new opportunities and benefits
  3. Big Data Management and Processing platforms that can handle Volume, Velocity and Veracity of the Data
  4. Understanding in-depth the architecture of Hadoop, HDFS, MapReduce and other Hadoop Ecosystem components and how them leverage Big Data solutions
  5. Understand and being able to apply Big Data Analytics lifecycle
  6. Being able to identify Big Data problems in cross-sectoral industry and determine right methodology, techniques and tools to solve these problems
- **Methods of instruction:** The class will be taught through lectures, including discussion around class examples/case studies, laboratory assignments and homework. Discussions based on student contributions add a vital and dynamic element to the class. Students are expected to come to the class with comments or questions from the course readings and actively participate in in-class discussions. Final project that will be assigned to student-groups of 2-3 students and their presentation will help to students to get experience of solving real data-driven problems of cross-sectoral business and share the experience they acquired to classmates.
- **Workload:** It is estimated that the students will need to spend 3-5 hours of study and preparation for the classes every week. Estimated amount of time to spend on course homework is additional 3-5 hours per week.

Materials

- **Primary or required books/readings for the course:**
  - Deepak Vohra, Practical Hadoop Ecosystem - A Definitive Guide to Hadoop-Related Frameworks and Tools, 2016, ISBN-13: 978-1-4842-2198-3, (electronic): 978-1-4842-2199-0
  - Dirk deRoos, Paul C. Zikopoulos, Bruce Brown, Rafael Coss, and Roman B. Melnyk, Hadoop for Dummies, 2014, ISBN: 978-1-118-60755-8

- **Supplemental or optional books/readings:**
  - Jared P. Lander, R for Everyone: Advanced Analytics and Graphics, 2014, ISBN-13: 978-0-321-88803-7
  - Wes McKinney, Python for Data Analytics, 2013, ISBN 978-1-449-31979-3
  - John White, Hadoop: The Definitive Guide 4th ed., 2015, ISBN 978-1-491-90163-2
  - Judith Hurwitz, Alan Nugent, Dr. Fern Halper, and Marcia Kaufman Judith Hurwitz, Big Data For Dummies, 2013, ISBN: 978-1-118-50422-2

- **Additionally some PDFs:**
  Additionally some PDFs from Open-source community members and vendors will be provided for further readings.

Requirements

- **Exams and quizzes:** Students will take 2 exams (midterm and final). These will be closed book (no books, no laptops or other devices) tests consisting of very limited number multiple-choice, open-ended test questions, problem solving using codding and understanding written code. Time and place will be communicated during the term.
- **Assignment/problem sets/projects/reports/research papers:** Homework(s) in a form of weekly/monthly written team assignments will be given during the term. These will be software development documents for a hypothetical project. Each homework assignment will be based on the previous assignments, reflecting subsequent phases of the project. Laboratory Assignments and Final Projects will be assigned to teams of 2-3 students. Detailed information and the exact dates will be communicated during the term. The students will submit the homework assignments online and in hard copy. The homework will be graded based on clarity, technical soundness, thoroughness and coverage, relevance to provided standards and utilization of resources.
- **Other requirement:** Academic honesty is required in all stages of exams, assignments, labs and projects.

Policies

- **Grading procedures:**
  - The students will be graded on absolute scale.

- The course grade will be calculated from the following components:
-

Midterm exam – 25%	Labs and quizzes – 20%	Final Project – 15%	Final exam – 35%	Attendance – 5%

- Students, who contend that their grade is not an accurate reflection of their accomplishments in the class, should first discuss their grade assessment with the instructor. For further steps please refer to the university procedures.

- **Attendance and tardiness:** Attendance is an indispensable element of the educational process. In compliance with Azerbaijani legislation, instructors are required to monitor attendance and inform the Registrar and the Dean of the respective School when students miss significant amounts of class time. Azerbaijani legislation mandates that students who fail to attend at least 75% of classes will fail the course.
  - ADA attendance policy excuses two (2) student absences, though these should reflect a serious need on the student's part to be away from class.
    In case of involuntary and unpredictable serious disruption of normal life, students may appeal to a grievance procedure through Office of the Dean of the School of Education.
- **Classroom decorum:** To avoid distractions late students are asked NOT to enter the class after the doors are closed. Cell phones shell be placed on silent mode or switched OFF, and shall NOT be used in the classroom during class sessions. As an exception, students may be allowed to leave or enter the room with the instructor's permission. *Students are not allowed to leave classroom to use their cell phones.*
- **Class participation:** Students are encouraged to contribute to class discussion. Certain percent of the course grade will depend upon contributions to class sessions. Class participation provides the opportunity to practice speaking and persuasive skills, as well as the ability to listen. What matters is the quality of one's contributions, not the number of times one speaks.
- **Missed or late assignments/extensions:** No missed assignments will be accepted later.
- **Standards for academic honesty and penalties for infractions:** If student found guilty of academic dishonesty first time, he or she would fail the course. If the case repeats again, student will be expelled. For more information please read the Honor Code.

Schedule

- **Tentative calendar of topics and readings:** The course is organized in 15 weeks.

Week of	Theme	Topics	Reading

1	Introduction to Big Data Analytics	General Information and class policy References and learning resources Digital Universe Volume and trends Why Data growth is so high? What is Big Data and When it becomes Big? Understanding 5Vs of Big Data Data concept and format of the available Data Big Data Landscape	Slides (Lecture content)
2	New Business Opportunities from Big Data Analytics (Use Cases)	Data Science and what a Data Scientist does Key skill-sets for Data Engineer and Data Scientist List Big Data and Data Science Use Cases Define Standard Parameters for Use Cases	book2-Ch1, book2-Ch2
3	Big Data Platforms, Hadoop Distributed File System (HDFS)	Understanding key differences of Big Data approach from Traditional IT approach Common Big Data Architecture Schema-on-Read vs. Schema-on-Write Why Big Data becomes Hot Topic right now? Introduction to Hadoop HDFS and MapReduce as Key Components of the Hadoop Understanding of architecture and working of HDFS and MapReduce	Book1-Ch1, PDF-tutorial
4	Setting Virtual Environment for Data Science	Installation and setting some of recommended tools for Data Scientist	Book1-Ch2 Book2-Ch3
	Quiz Exam 1:	All Covered Topics	
5		Key components of Hadoop Ecosystem	

	Hadoop Ecosystem (Primary components: Hadoop, YARN, MapReduce, HDFS, Pig, Hive, HBase, Zookeeper, Sqoop, Flume)	Relation between Components Classification of Hadoop Ecosystem's Components according to assigned Duties Core Hadoop's Components Data Access Data Storage Data Integration	Slides, PDF-tutorial
6	Hadoop Installation and Configuration (Laboratory)		Stet-by-step guidelines will be provided
	Laboratory Assignment 1:	Setting Hadoop Platform	
7	Data Mining Techniques and Tools		
8	**Midterm exam** Distributed Computing with Hadoop and MapReduce (YARN)	MapReduce Programming Concept	
9	**Spring Break**		
10	Brief introduction to programming in R	Installation, R Studio IDE, R packages for Data Management and Analytics. R and Python – Which one to use?	R for Everyone, chapters 1-3; Python for Data Analytics chapters 3, 4
	Homework	R programming basics	

11	Data Manipulation and Processing with R and Python	Standard Structures available in R and Python. Advanced Data Structures.	R for Everyone, chapters 4,5,6; Python for Data Analytics chapters 6, 7
	Quiz Exam 2:	Understanding and being able to use Data Structures available in R. Data manipulation with R.	R for Everyone, chapters 15,16
12	Big Data Statistical Analytics and ML Algorithms		Resources will be provided
13	Text Analytics		Resources will be provided
	Laboratory Assignment 2:	Using Hadoop Ecosystem components for Data Analytics	
14	**Final Projects Presentations**	Final project presentations / Presentation	
15	Big Data Security and Privacy		Final project presentations / Presentation
	Final exam	Final exam	Final exam

- **Last day to withdraw from the course:** March 31, 2017

Resources

- **Support services on campus:**
  - Students are encouraged to consult with the Writing Center for checking their papers and assignments. Please visit the Writing Center or contact them by email: writingcenter@ada.edu.az
  - Adjusting to student life, pursuing academic and personal goals can be emotionally stressful and challenging. Students are encouraged to make individual appointments with Counselor to receive a professional psychological support. Please contact Zohra Malikova, Counselor at phone (012) 4373235 x164 or by email: zmelikova@ada.edu.az
- **Tips for success**
  - Students will need to read the course readings throughout the term to learn the material and to be able to contribute to class discussions.

- Here are some words of wisdom from *Dr. Morgan Liu* of The Ohio State University on "How to Read an Academic Book or Article": Reading an academic article/book is not like reading a newspaper or novel. Following these guidelines will help keep you from being overwhelmed, and make you better prepared for discussions & essays.
  1. Read actively, not passively. You read because you are trying to mine the text for insights. You are not reading because you have to get through it. Take an active posture while reading: you are trying to take something away from the reading.
  2. Before you begin, ask yourself: what is my purpose for reading this? First ask yourself: What topic is the course covering this week? What are the active issues and recurrent themes? What sorts of insights do I hope to get out of the reading? The Reading Questions will help you get a grip.
  3. Do not always read from start to finish. Read the introduction or opening paragraphs. Then skip to the back and read the conclusion to see where the thing is going. Flip through the article/book and take note of the section or chapter titles. Read the beginning & end of each section to see what they're about. Stop. Think about what this article/book is trying to accomplish and how it will get there. Get a sense of the overall arguments first, and how the author will develop them. Then step back, close your eyes and think, what are the most important parts that I must read? What can I skim over for now?
  4. Read selectively. Do not read every word in the text. Read the most important parts first, and see what else you need to read as you go. You can always go back. You have my permission to skip the less important parts – no guilt, really!! But you got to be thoughtful to figure what those are. Better to read the most important parts thoughtfully, than try to get through the entire thing like a zombie.
  5. Stop frequently and ask yourself: what did I just learn? Make notes as you go. Write down questions. Don't get bogged down in unimportant detail. If your mind starts to wander, stop and refocus on the big picture: what's been happening in the text, and where is it.

Statement on Accommodation

- ADA University provides upon request appropriate academic accommodations for qualified students with documented disabilities. Any student who feels she/he may need an accommodation based on the impact of a disability should notify the Office of Disability Services and Inclusive Education about his/her needs before the start of the academic term. Please contact Elnur Eyvazov, Director of the Office of Disability Services and Inclusive Education at phone: (012) 4373235 x249 or by e-mail: eeyvazov@ada.edu.az
- Reasonable accommodation is possible for students' religious beliefs, observations, and practices or for foreseeable conflicts because of athletic competition

Disclaimer

- **This syllabus, including the course schedule is subject to change as necessary and students will be notified accordingly.**