**BUAN (Big Data) Course Syllabus**
**Spring 2020**
**(Content and dates are subject to changes)**

**Course Information**

**Course Number:** BUAN 6346 Sec 502
**Title**: Big Data
**Term:** Spring 2020
**Class Hours**: Thursday 7:00pm-9:45pm
**Location:** Webex conference. Email calendar invite has been sent. Link is posted on eLearning

**Instructor Information**

**Name**: Uri Smashnov
**Email**: uri.smashnov@utdallas.edu
**Office Hours**: Via Webex conference, posted on eLearning. Thursday 5-6pm. Additional sessions by email request.

**Teaching Assistant:**

**Name:** Sheraaz Soheil Ahmed
**Email:** SheraazSoheilAhmed@UTDallas.edu
**Office hours:** Wednesday 4pm to 5pm. Via Webex, the link will be posted on eLearning

**Course Description**

The course covers Big Data concepts, architecture, and hands-on use of several tools in the Hadoop Ecosystem. The course will cover the theoretical as well as hands-on. The tools covered include Linux, Hadoop, Sqoop, Flume, Hive, Hbase, and Spark (2.2.1).
Overview of Cloud providers and their offerings in the Big Data domain.

**Primary Learning Objectives**

- Know the basics of Linux, Hadoop, HDFS, Hive, Hbase, Sqoop, and Flume
- Be able to successfully ingest and manipulate data in Hive
- Understand the basics of Spark and being able to use Spark (Python API) for data manipulation

**Hardware and Software Requirements:**

- 64 bit OS, Windows Laptop with **Solid State Drive (SSD)**, with 50 Gb of free space and at least 8GB of memory. Hadoop/Linux sandbox requires at least 4 Gb of memory to run correctly. (Tested only with Windows 10 – 64 bit).
- You must bring your laptop to every class and exam
- Mac users can install Windows as your secondary OS. UT Dallas has Windows OS licenses available for students via https://www.utdallas.edu/oit/howto/imagine/ . If you prefer to use Mac OS, licensed VMware Fusion Pro 11 is available through VMware academic program for $95 (at least one student had success with such set up).
- It is not possible to run VMware software in already virtualized environment. It means that one cannot run it on Azure or AWS type of environment.

**Software Requirements:**

- As a base, we will be utilizing Cloudera 5.4.3 sandbox with Hadoop and Tools pre-installed. The sandbox will be available for download before the class starts.
- We will use more advanced versions of several tools than provided by Cloudera, for example we will be using Spark 2.2.1 to take full advantage of DataFrames and SQL.

**Textbooks:**

***Linux:*** The Linux Command-line: A Complete Introduction, 1st Edition (January 2012) by William E. Shotts, Jr.

http://linuxcommand.org/tlcl.php
https://www.nostarch.com/tlcl
https://www.amazon.com/Linux-Command-Line-Complete-Introduction/dp/1593273894

*Safary Books Playlist: https://learning.oreilly.com/playlists/8d86ce15-205d-488b-a0a3-3f6509dc6b46*

***Hadoop:*** Hadoop: The Definitive Guide, 4th Edition (March 2015)

O'Reilly Publishing

*This will teach you 70% of everything you ever need to know about Hadoop and MapReduce. This book has separate chapters dedicated to Hive, Kafka, Sqoop, and Flume that are also very good.*

***Python:*** How to Think Like a Computer Scientist

*http://openbookproject.net/thinkcs/python/english3e/*

***Spark:*** Spark: The Definitive Guide. Big Data Processing Made Simple. By Bill Chambers and Matei Zaharia.

**Communication:**

- We will be using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from classmates, the TA, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza.

- I'm testing MS Teams as an alternative to Piazza.

- For private questions, please use UTD email.

**Other Course Materials**

**Scores and Grade**: Assignment, project, and exam scores as well as the final grade will be available at the UTD eLearning site.

**Class Notes**: Class notes will be available through UTD eLearning site

**Individual Assignments**: There will be several assignments and labs.

- *Individual assignments must be done by individual students.* You may discuss assignments with pears, however all work should be done individually.
- No hand-written project report will be accepted; it must be word-processed or type-written.
- Submission of assignments or projects via e-mail, fax, or other electronic media is not acceptable.

**Individual Project**: project details will be announced in the third week of classes. You are expected to start working on the project during semester

**Exams**

- There will be three in-class exams.
- Students will use their Sandbox to answer exam questions.
- Exams will be "open book", with all materials allowed.
- Each student should open Webex account. Webex will be used to record students screen while performing exams. The recording will be shared with the professor after the exam. https://www.utdallas.edu/oit/howto/video-conferencing/webex/

**Letter grades** will be assigned based on the following ranges using total class score.

| Grade | Start | End |
|---|---|---|
| A | 91 | 100 |
| A- | 87 | 91 |
| B+ | 83 | 87 |
| B | 79 | 83 |
| B- | 75 | 79 |
| C+ | 70 | 75 |
| C | 60 | 70 |
| F | 0 | 60 |

Calculated Grade weights:

- Exam 1 — (15%)
- Exam 2 — (20%)

- Exam 3 — (20%)
- Homework (Labs and Assignments) — (20%)
- Semester Project — (25%)

Calculated Grade might include “curve”.

## Course Policies

- The in-class exercises are to be completed in class and are due by the end of class, unless otherwise stated in eLearning. You must be present in the class to complete in-class exercises.
- You must be able to self-support yourself:
    - Install VMware Player
    - Launch provided Sandbox
    - Backup your Sandbox or individual scripts as needed and at least once a week
    - Start/Stop Hadoop tools as needed using provided scripts and instructions
- If you are not familiar with Linux, it is strongly recommended to go over parts 1~6. You can use terminal in the provided Sandbox to practice Linux commands.
    - http://linuxcommand.org/tlcl.php
    - http://linuxcommand.org/lc3_learning_the_shell.php

- There will be at least 7 days allocated for an assignment. You need to start working on the assignments as soon as assignment is posted and identify any potential difficulties with input files, and/or provided scripts. There will be no extensions due to technical difficulties.
- You must back up all your files. This includes your programs and your data. You can use VMware Player “shared folder” to copy your files to local machine. It is strongly recommended to store/copy of zipped (7-zip is reliable tool for very large files) Sandbox on UTD Box, or any other storage independent of your Laptop. The policy applies if you use AWS/Azure as well.
- Assignments must be submitted through eLearning. Emailed submissions are not accepted.
- There will be 6 hours grace period before penalty starts in case of late assignments. Late assignments are accepted for up to 3 days - with a 25% points deduction. After 3 days there will be 100-point deduction. Not every assignment will have late submission option due to proximity to exam and need to publish the solution.
- Exams will contain significant portion of hands-on work. You'll be given a problems to solve, during the exam.
- Students are expected to monitor Piazza and eLearning daily.
- Missed exams earn a 0/100.
- A signed note from a medical doctor will be required for any grading impacted policy. This must include the physician’s name and contact information for verification.
- **Makeup Exams**: There will be no make-up exams, except for medical emergency (written statement justifying the situation from a physician required) or other university accepted reason.
- **Cheating will not be tolerated.** When I find evidence of cheating, the documentation is turned over to the Office of Community Standards and Conduct.
- **All work is individual work**
- Read on academic dishonesty here: https://www.utdallas.edu/conduct/dishonesty/

**Class Calendar (all dates are subject to changes):**

| | Date | Topic/Lecture | Reading (prior to Lecture) and Notes |
|---|---|---|---|
| 1 | 1/16 | Course Introduction<br>Syllabus review<br>Sandbox and Tools Review | |
| 2 | 1/23 | Linux & Hadoop<br>Hands-on: Linux shell and HDFS | Read Chapters 1-6, in The Linux Command-line by William E. Shotts, Jr. |
| 3 | 1/30 | Hadoop Architecture, data storage and ecosystem<br>Hands-on: HDFS, submitting MapReduce jobs, review Logs, Hue | Read Chapters 1, 3, 4, and 7 in Hadoop: The Definitive Guide, 4th Edition. In Chapter 2, read "Analyzing the Data with Hadoop" starting at bottom of p22 and ending at top of p24. |
| 4 | 2/6 | Sqoop and Flume<br>Hands-on: Sqoop and Flume | Read Chapters 14 and 15 in Hadoop: The Definitive Guide, 4th Edition |
| 5 | 2/13 | Hive<br>Hands-on: Hive | Read Chapters 17 in Hadoop: The Definitive Guide, 4th Edition |
| 6 | 2/20 | Exam 1 | In class on personal Laptop (All materials prior to Hive, Hive is not included) |
| 7 | 2/27 | Hive<br>Hands-on: Hive | |
| 8 | 3/5 | Hbase<br>Hands-on: Hbase | Read Chapters 20 in Hadoop: The Definitive Guide, 4th Edition |
| 9 | 3/12 | Exam 2 | In class on personal Laptop |
| | 3/19 | Spring Break | Spring Break |
| 10 | 3/26 | No Class due to University policy | |
| 11 | 4/2 | • Python and PySpark Spark Architecture. Jupyter Notebook for PySpark interface.<br>• Big Data and Cloud Providers offerings | http://openbookproject.net/thinkcs/python/english3e/<br>Read Chapters 2 & 3 in Spark*: The Definitive Guide* |
| 12 | 4/9 | Spark 1 | Read Chapters 2,3,4,5 in Spark*: The Definitive Guide* |
| 13 | 4/16 | Spark 2 | Read Chapters 6,7 & 8, 10 in Spark*: The Definitive Guide* |
| 14 | 4/23 | Spark 3 | Read Chapters 16 in Spark*: The Definitive Guide* |
| 15 | 4/30 | **Exam 3** | Remote during class hours. On Personal Laptop |

**UT Dallas Syllabus Policies and Procedures**

The information contained in the following link constitutes the University's policies and procedures segment of the course syllabus.

Please go to **http://go.utdallas.edu/syllabus-policies** for these policies.

***The descriptions and timelines contained in this syllabus are subject to change at the discretion of the Professor.***

**Accommodations**

It is the policy and practice of The University of Texas at Dallas to make reasonable accommodations for students with properly documented disabilities. However, written notification from the Office of Student Accessibility (OSA) is required. If you are eligible to receive an accommodation and would like to request it for this course, please discuss it with me and allow one week advance notice. Students who have questions about receiving accommodations, or those who have, or think they may have, a disability (mobility, sensory, health, psychological, learning, etc.) are invited to contact the Office of Student Accessibility for a confidential discussion. OSA is located in the Student Services Building, suite 3.200. They can be reached by phone at (972) 883-2098, or by email at studentaccess@utdallas.edu.