NATIONAL UNIVERSITY
of Computer & Emerging Sciences, Lahore

Department of Data Science

FUNDAMENTALS OF BIG DATA ANALYTICS

SPRING 2022

Instructor Name: Dr. Iqra Safder **TA Name (if any):**
Email address: iqra.safder@nu.edu.pk **Email address:**
Office Location/Number: C-135 **Office Location/Number:**
Office Hours: Tuesday 1:00-3:00 PM, Thursday 1:00pm to 3:00pm

Course Information

Program: BSDS **Credit Hours:** 3 **Type:** Core

Pre-requisites (if any): Programming competence, Intro to Data Science

Course Website (if any) : Google classroom
Class Meeting Time:

Course Description/Objectives/Goals:

With the proliferation of unstructured data in quantities that impede the use of traditional statistical approaches. To examine such data, new techniques are required. In order to be responsive, new algorithms are required to deal with different techniques. New data storage and retrieval mechanisms are required. Many of the algorithms come from well-known big data owners like Google (search, adwords), Amazon (recommended books), and Facebook (social network analysis). As more players enter the competition, new methods will emerge to meet their needs. The course objective is to develop understanding about the core concept of Big Data, why Big Data requires a different programming paradigm and mindset, and what are the various programming approaches used, what type of data can be processed.

Course Learning Outcomes (CLOs):		
At the end of the course students will be able to:	**Domain**	**BT* Level**
Understand the fundamental concepts of Big Data and its programming paradigm.		C2
Prepare and wrangle the data for analysis,		C2
Hadoop/MapReduce Programming, Framework, and Ecosystem		C3
Apache Spark Programming		C3
* BT= Bloom's Taxonomy, C=Cognitive domain, P=Psychomotor domain, A= Affective domain.		

Bloom's taxonomy Levels: 1. Knowledge, 2. Comprehension, 3. Application, 4. Analysis, 5. Synthesis, 6. Evaluation

Textbook(s) /Supplementary Readings:

There is no standard one "textbook" for this course. The following book will be used as a primary text to guide some of the discussions, but it will be heavily supplemented with lecture notes and reading assignments from other sources.

Mining of Massive Datasets by Jure Leskovec Stanford Univ. Anand Rajaraman Milliway Labs , Jeffrey D. Ullman Stanford Univ.

Additional references and books related to the course:

- White, Tom. “Hadoop: The definitive guide." O'Reilly Media, Inc., 2012. 2.
- Karau, Holden, Andy Konwinski, Patrick Wendell, and Matei Zaharia. “Learning spark: lightning-fast big data analysis." O'Reilly Media, Inc., 2015. 3.
- Miner, Donald, and Adam Shook. “MapReduce design patterns: building effective algorithms and analytics for Hadoop and other systems." O'Reilly Media, Inc., 2012.

Tentative Weekly Schedule

Week	Topics to be covered	Readings	Assignments /Projects?
1	• What is Big Data and Why we need Big data • Sources of Data • Intro to big Data tools, • Data Engineering • Types of Data		

2	• Exploratory Data Analysis
	  ○ Data Objects
	  ○ Types of Attributes
	  ○ Basic tools (plots, graphs and summary statistics) of EDA,
	  ○ Intro to Techniques for Data Analytics (classification, Clustering, regression)		

3	• Clustering and Cluster Analysis
	  ○ Partitional Clustering
	  ○ Hierarchical Clustering
	  ○ Density based Clustering
	•		

4-5	• Hadoop as a Platform, Hadoop Distributed File Systems (HDFS), • MapReduce Framework,		
6-7	• Apache Scala Basic, Apache Scala Advances		
8- 9	• Resource Management in the cluster (YARN)		
10-11	• Resilient Distributed Datasets (RDD)		
11-12	• Apache Spark, Apache Spark SQL		
12-13	• Machine learning on Hadoop / Spark		
14-15	• Spark Streaming, Other Components of Hadoop Ecosystem		

(Tentative) Grading Criteria

Quizzes	10%
Assignments/Homeworks/Project	*15* - 25%
Midterms	25-30%
Final Exam	40 - 45%
Total:	**100 %**

Course Policies

- *Course outline may change 10-20% as we proceed in the semester. We may add and remove a few topics.*
- ***Grading scheme: Relative***
- Depending on the situation of COVID 19, this weightage of midterms can be reduced and added in assignments/homeworks/project.
- *Weightage of other evaluations can also be adjusted if needed.*
- *Assignment deadlines for assignment and Project are hard.*
- *NO Cell Phone usage in class, they must be turned off at all times.*
- *There will be no retake of quizzes or exams.*
- ***Integrity in the assignments/quizzes is expected; otherwise result would be an F grade in the course or may be the case is forwarded to Disciplinary committee.***
- *Attendance MUST be ensured according to the University policy to avoid disqualification.*