University of Central Punjab

Faculty of Information Technology

Course Code	CSDS4423
Course Title	Programming for Big Data
Credit Hours	3
Prerequisites by Course(s) and Topics	CSCP2023 & CSCP2021
Assessment Instruments with Weights *(homework, quizzes, midterms, final, programming assignments, lab work, etc.)*	Quizzes (10%) + Assignments + Tutorials (10%), Mid Term (20%), Final Term (45%), Project/Presentations (10%) and Class Participation (5%)
Semester	SPRING 2021
Course Instructor	Saeed Iqbal Khattak
Course Coordinator	Dr. Adnan N. Qureshi
Lab Instructor	
Office Hours	To be announced later.
Plagiarism Policy	If anybody found in act of plagiarism, he/she will be marked zero in all of his/her instruments of that category. Plagiarism offense in midterm and final term will result in (F) grade. Marks will be uploaded on portal and can be contested within a week or would be considered final.
Current Catalog Description	The course aims to introduce basic concepts that will help students to: 1. Understanding of Python 2. Data manipulation using python 3. Hands on practice with python libraries (Numpy, Pandas and Matplot). 4. Data Visualization 5. Introduction Big Data 6. Introduction to Spark 7. Introduction to RDDs 8. Introduction to DataFrames 9. Advanced Spark Topics 10. Introduction to Spark MLlib
Textbook (or Laboratory Manual for Laboratory Courses)	1. Mining of Massive Datasets 2. Data Analysis with open source tools 3. Learning Apache Spark with Python 4. Python for Data Analysis 5. Python for Programmers 6. Big Data, Mining, and Analytics
Reference Material	https://saeediqbal.github.com
Course Goals	This course is for students who have some programming and database experience. The objective of this course is to give students some experience in data analysis and developing applications that utilize the vast amount of data that is available to general public to create programs that provides information used in improving the standard of application performance. Discovering how the efficiency of applications can be improved by understanding the data.

<table>
<tr><td>Learning Outcomes</td><td>1. Learn basic concepts required for big data applications
2. Design and develop algorithms to collect and present data into an information.
3. Develop algorithms to display real-time content using Distributed Programming.
4. Implement interactive real-time distributed framework.
5. Configured advanced real-time applications frameworks.
6. Learn, implement and integration of web services with big data applications.</td></tr>
<tr><td>Topics Covered in the Course, with Number of Lectures on Each Topic (assume 15-week instruction and 1.5 hour lectures)</td><td>Attached</td></tr>
<tr><td>Programming Assignments Done in the Course</td><td>Yes</td></tr>
<tr><td>Exam:</td><td>Mid : 1.5Hrs
Final : 3Hrs</td></tr>
<tr><td>Class Time Spent on (in credit hours)</td><td><table><tr><th>Theory</th><th>Problem Analysis</th><th>Solution Design</th><th>Social and Ethical Issues</th></tr><tr><td>0.5</td><td>1</td><td>1</td><td>0.5</td></tr></table></td></tr>
<tr><td>Oral and Written Communications</td><td></td></tr>
</table>

Week	Topics Covered	Instruments
1	• Discussion on Python and its market position. • Motivation regarding learning aspects of this course • Setting up environment for Python. • Installation of Anaconda • What is Data? • What is Big Data? • Characteristics of Big Data • What are the Vs of Big Data? • The Impact of Big Data • Big Data - Beyond the Hype, Big Data Examples, Sources of Big Data • Big Data Adoption, The Big Data and Data Science • The Big Data Platform, Big Data and Data Science. Skills for Data Scientists	
2	• Machine-Generated Data: People Generated Data , Organization generated data • Characteristics of Big Data types , volume , velocity , variety, veracity ,value • Building a Big Data Strategy , Component of big data • Types of IDE(s) and WIDE that will be used in the duration of this course. e.g. Spyder, Jupyter etc • Hello World Program "Print Command" • Keyword Types • Expressions and Variables	Assignment 1
3	• Input Method • Conditions and Branching • Loops • String Operations • Lists and Tuples • Sets • Dictionaries	Quiz 1
4	• Data Analysis Process • Steps of processes: acquiring, exploring, pre-processing, Analyzing, communicating and turning into action • What is a Distributed File System? • Scale-able Computing over the Internet, Programming Models for Big Data • Reading and Writing files • Functions • Objects and Classes	
5	• Working with Pandas • Descriptive Statistics with Pandas • Group by with Python • Data Manipulation with Pandas	Assignment 2
6	• Data Wrangling with Pandas • Data Manipulation with Pandas	Quiz 2
7	• Introduction with Numpy • Numpy one dimensional Array • Numpy two dimensional Array • Numpy Array Operations	Quiz 3

8	• Introduction to Matplotlib • Basic Plotting with Matplotlib • Line Plots • Area Plots • Histograms • Bar Charts • Pie Charts • Box Plots • Scatter Plots • Word Cloud	
	Revision	
	MID TERM	
9	• Introduction to Hadoop • Hadoop: Why, Where and Who? • The Hadoop Ecosystem: Welcome to the zoo! • The Hadoop Distributed File System: A Storage System for Big Data • YARN: A Resource Manager for Hadoop	Review
10	• The Hadoop Distributed File System: A Storage System for Big Data • YARN: A Resource Manager for Hadoop • What is Spark and what is its purpose?	Assignment 3
11	• Components of the Spark unified stack • Resilient Distributed Dataset (RDD) • What is Pig and Hive • Architecture of Pig and Hive	Quiz 4
12	• Understand how to create parallelized collections and external datasets • Work with Resilient Distributed Dataset (RDD) operations • Utilize shared variables and key-value pairs	Assignment 4
13	• Describe and run some Spark examples • Pass functions to Spark • Create and run a Spark standalone application	Quiz 5
14	• Introduction to Apache Kafka • Components of Apache Kafka • Internal Architecture of Apache Kafka	Assignment 5 and Quiz 6
15	• Introduction to Apache Zookeeper • Components of Apache Zookeeper • Internal Architecture of Apache Zookeeper	Quiz 7

Final Term Project

1. Students will create a final project on a topic of their choice using the technologies and techniques covered in this course.
2. Student will create a short (less than 10 minute) presentation explaining how to use their project and where requirement were met.
3. Final project presentations will take place during the week of finals. Exact date TBA.