**University of Central Punjab**

**Faculty of Information Technology**



**BSCS**

**PROGRAM (S) TO BE**

**EVALUATED**

1. **Course Description**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Course Code** | CSDS4423 | | | |
| **Course Title** | Programming for Big Data | | | |
| **Credit Hours** | 3 | | | |
| **Prerequisites by Course(s) and Topics** | Object oriented programming | | | |
| **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 2024 | | | |
| **Course Instructor (s)** | Misbah naz | | | |
| **Course Instructor Email** | [Misbah.naz@ucp.edu.pk](mailto:Misbah.naz@ucp.edu.pk) | | | |
| **Course Coordinator** | Dr Saeed Iqbal khattak | | | |
| **Office Hours** | Tuesday 12-1pm  Thursday 10-11 pm  Friday 12.05 -1pm | | | |
| **Office Location** | A- F202 Cabin 2 | | | |
| **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. | | | |
| **Tools Used in the Course** | Hadoop, Yarn, Spyder, Anaconda | | | |
| **Course Description** | 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 perfor- mance. Discovering how the efficiency of applications can be improved by understanding the data. | | | |
| **Course Objectives** | 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   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** |  | | | |
| **Programming Assignments Done in the Course** | Yes | | | |
| **Class Time Spent on** (in credit hours) |  |  |  |  |
|  |  |  |  |
| **Oral and Written Communications** |  | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **CLO** | **CLO STATEMENT** | **Bloom’s Taxonomy Level** | **PLO** |
| 1 | Comprehend the fundamental constructs of Big Data. | C2 (Understand) | 2 |
| 2 | Comprehend and implement the fundamental constructs of Distributed computing tools. | C2 (Understand) | 3 |
| 3 | Apply various libraries for plotting, interpreting and  analyzing data in Python. | C3 (Apply) | 5 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Topics** | **Book / Slides Reference** | **Evaluation Used** | **Relation to CLO** |
| **1st** | * 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 |  |  | CLO 1 |
| **2nd** | * 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 |  | CLO 1 |
| **3rd** | * Input Method * Conditions and Branching * Loops * String Operations * Lists and Tuples * Sets   Dictionaries | Quiz 1 |  | CLO 1 |
| **4th** | * Data Analysis Process * Steps of processes: acquiring, exploring, pre-processing, Analyzing, com- municating 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 |  |  | CLO 2 |
| **5th** | * Working with Pandas * Descriptive Statistics with Pandas * Group by with Python   Data Manipulation with Pandas |  | Assignment 2 | CLO 2 |
| **6th** | * Data Wrangling with Pandas   Data Manipulation with Pandas |  | Quiz 2 | CLO 2 |
| **7th** | * Introduction with Numpy * Numpy one dimensional Array * Numpy two dimensional Array   Numpy Array Operations |  | Quiz 3 | CLO 2 |
| **8th** | * Introduction to Matplotlib * Basic Plotting with Matplotlib * Line Plots * Area Plots * Histograms * Bar Charts * Pie Charts * Box Plots * Scatter Plots   Word Cloud |  |  | CLO 2 |
|  | **MID Term**  **MID Term Review** |  |  | CLO 1, CLO2 |
| **9th** | * 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 | CLO 3 |
| **10th** | * 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 | CLO 3 |
| **11th** | * Components of the Spark unified stack * Resilient Distributed Dataset (RDD) * What is Pig and Hive   Architecture of Pig and Hive |  | Quiz 4 | CLO 3 |
| **12th** | * 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 | CLO 1, CLO 3 |
| **13th** | * Describe and run some Spark examples * Pass functions to Spark   Create and run a Spark standalone application |  | Quiz 5 | CLO 3 |
| **14th** | * Introduction to Apache Kafka * Components of Apache Kafka   Internal Architecture of Apache Kafka |  |  |  |
| **15th** | * Introduction to Apache Zookeeper * Components of Apache Zookeeper   Internal Architecture of Apache Zookeeper |  |  |  |
| **16th** | * **Revision** |  |  | CLO\_1, CLO\_3 |
|