AMRITA VISHWA VIDYAPEETHAM

School of Computing

(AMARAVATI, AMRITAPURI, BANGALORE, COIMBATORE, CHENNAI)

B. TECH Minor in Data Science

CURRICULUM 2022

GENERAL INFORMATION

ABBREVIATIONS USED IN THE CURRICULUM

Cat	-	Category
L	-	Lecture
T	-	Tutorial
P	-	Practical
Cr	-	Credits
ENGG	-	Engineering Sciences (including General, Core and Electives)
MAT	-	Mathematics
CGPA	-	Cumulative Grade Point Average

Course Outcome (CO) – Statements that describe what students are expected to know, and are able to do at the end of each course. These relate to the skills, knowledge and behaviour that students acquire in their progress through the course.

Program Outcomes (POs) – Program Outcomes are statements that describe what students are expected to know and be able to do upon graduating from the Program. These relate to the skills, knowledge, attitude and behaviour that students acquire through the program. NBA has defined the Program Outcomes for each discipline.

PROGRAM OUTCOMES FOR ENGINEERING

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

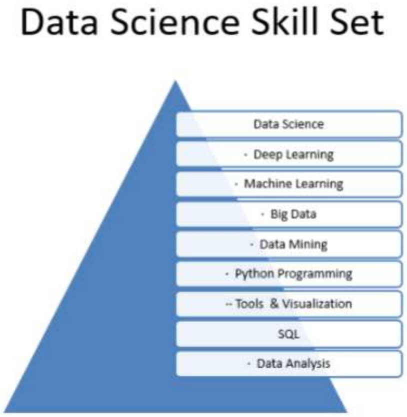
MOTIVATION:

With the proliferation of sensors and devices in Internet of Everything(IoE), modern businesses are awash with data, big data to be precise. There is a vast demand for professionals who are able to unlock the power of big data using cutting-edge technology to obtain actionable insights.

Data Science is the study of data. It is about extracting, analyzing, visualizing, managing and storing data to create insights. These insights help the companies to make powerful data-driven decisions. Data Science requires the usage of both unstructured and structured data. It is becoming clear by the day that there is enormous value in data processing and analysis—and that is where a data scientist steps in adding value to business:

- Empowering management to make better decision
- Directing actions based on trends—which in turn help to define goals
- Challenging the staff to adopt best practices and focus on issues that matter
- Identifying opportunities
- Decision making with quantifiable, data-driven evidence

Data Science being a multidisciplinary field that has its roots in statistics, math and computer science, posits immense value to students across disciplines in their multi-fold application areas. The various skill sets required in this domain is illustrated below:



A Minor in Data Science, therefore, brings in this added advantage to a graduating student.

PROGRAM DESCRIPTION - PROGRAM SPECIFIC OUTCOMES FOR DATA SCIENCE MINOR

On completion of a Minor in Data science, a student will be able to:

PSO1: Conduct exploratory data analysis creating visualizations, identify patterns and employ machine learning algorithms to derive insights to make powerful data-driven decisions.

PSO2: Apply statistical and mathematical concepts to solve computational tasks employing inferential thinking to model real world problems abound with big data.

PREREQUISITES FOR PURSUING DATA SCIENCE MINOR

Students who have a CGPA of 7.5 or above at the end of their second semester are eligible to register for this minor.

Cat.	Code	Title	L T P	Credit
ENGG	23CSE231M	Introduction to Data Science and Analytics	2 0 3	3
ENGG	23CSE232M	Python for Data Science	2 0 3	3
ENGG	23CSE233M	Database Management Systems for Data Science	2 0 3	3
ENGG	23CSE234M	Data Visualization	2 0 3	3
ENGG	23CSE235M	Machine Learning with Python	2 0 3	3
ENGG	23CSE236M	Big Data Frameworks for data science	2 0 3	3

23CSE231M Introduction to Data Science and Analytics L-T-P-C: 2-0-3-3

Pre-Requisite(s): Nil

Course objectives

This course introduces the scope of data science and analytics. Statistical fundamentals required for data science are introduced. Overview of tools for data science is given. Data science project life cycle is discussed. Exploratory Data Analysis and the Data Science Process are illustrated.

Course Outcomes

After completing this course, the students will be able to

CO1: Understand and describe the role of data science and its tools.
CO2: Understand and describe the role of big data and cloud computing in data science.
CO3: Apply mathematical and statistical principles to the analysis of data.
CO4: Apply the techniques of Exploratory Data Analysis.
CO5: Apply correlations, distributions and hypothesis tests for inference.

CO-PO Mapping

PO/PSO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO														
CO1	2	1	2	2	3	2		1		1		2	2	2
CO2	2	1	2	2	3	2		1		1		2	2	3
CO3	3	3	3	3	3			2	1	1		2	3	3
CO4	3	3	3	3	3			2	1	1		2	3	2
CO5	3	3	3	3	3			2	1	1		2	3	3

Syllabus

Unit 1

Introduction - Overview of Data Science – Data Science roles – Career paths – Applications – Types of Analytics -Big data and its role in Data Science – Overview of Big data frameworks – Data science and cloud computing – role of cloud – cloud infrastructure - Essential Statistics for Data Science: Sampling, Sample Means and Sample Sizes - Descriptive statistics: Central tendency, dispersion, variance, covariance, kurtosis, five point summary.

Unit 2

Introduction and overview of Data Science tools – Python, R, SQL – Data science project life cycle - Data Pre-processing: Data cleaning, Data reduction, Data transformation, Data discretization - Datasets and their role in analytics – EDA - Role of Visualization and Graphing – Introduction to Visualization tools

Unit 3

Exploratory Data Analysis and the Data Science Process - Basic tools (plots, graphs and summary statistics) of EDA - Philosophy of EDA - The Data Science Process – Correlation - Randomness and Probability – Distributions – Hypothesis Test and inference

Text Book(s) / Reference(s)

Ani Adhikari. John DeNero, David Wagner, Computational and Inferential Thinking: The Foundations of Data Science, 2nd Edition, GitBook, 2019.

Shmueli G, Bruce PC, Yahav I, Patel NR, Lichtendahl Jr KC. Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons; 2018.

Schutt R, O'Neil C. Doing data science: Straight talk from the frontline. First Edition, O'Reilly Media, Inc.; 2013

Foster Provost and Tom Fawcett, Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. 2013

Avrim Blum, John Hopcroft and Ravindran Kannan, Foundations of Data Science, Cambridge University press, 2020

Wes McKiney, Python for Data Analysis, Data wrangling with Pandas, Numpy and Jupyter, O'Reilly, 3rd Edition, 2022

Evaluation Pattern: 70:30

Assessment	Internal	End Semester
Midterm	20	-
*Continuous Assessment (Theory) (CAT)	10	-
Continuous Assessment (Lab) (CAL)	40	-
End Semester	–	30

•CA – Can be Quizzes, Assignment, Projects, and Reports

23CSE232M	Python for Data Science	L-T-P-C: 2-0-3-3

Pre-Requisite(s): Nil

Course objectives

This course introduces python programming. Relevance of data structures in python is discussed. NumPy and its applications are discussed. Role of Pandas is illustrated towards data manipulation. The techniques to deal with data sets, their creation and management are introduced. Data wrangling is introduced. Relevant background required for dealing time series data is introduced.

Course Outcomes

After completing this course, the students will be able to

CO1: Apply Python for data science applications using the relevant data structures.
CO2: Apply NumPy libraries for applications involving array processing.
CO3: Apply Pandas libraries for data manipulation and Aggregations.
CO4: Apply libraries for time series applications.
CO5: Apply the techniques of data wrangling.

CO-PO Mapping

PO/PSO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO														
CO1	3	3	3	2	3			2	1	1		2	2	2
CO2	3	3	3	2	3			2	1	1		2	1	2
CO3	3	3	3	3	3	1		2	1	1		2	2	3
CO4	3	3	3	3	3	1		2	1	1		2	2	3
CO5	3	3	3	3	3	1		2	1	1		2	2	3

Syllabus

Unit 1

Python – Data types – Data structures in python – Programming with python – loading, reading and writing files in python – saving data - Lists, tuples, dictionaries and sets - lambda inline functions.

NumPy - Understanding Data Types in Python - NumPy Arrays - Array Indexing - Array Slicing: Accessing Sub arrays -Multidimensional sub arrays - Reshaping of Arrays -Array Concatenation and Splitting - Computation on NumPy Arrays - Array arithmetic – Aggregations -Computation on Arrays: Broadcasting - Working with Boolean Arrays - Sorting Arrays – Indexing – Binning Data

Unit 2

Data Manipulation with Pandas - Pandas Objects - Pandas Series Object - Series as specialized dictionary - Pandas DataFrame Object - Constructing DataFrame objects - Pandas Index Object - Data Indexing and Selection - Indexers: loc, iloc, and ix - Data Selection in DataFrame - Operating on Data in Pandas - Handling Missing Data –

Hierarchical Indexing - Methods of MultiIndex Creation - Multi-Indices - Data Aggregations on Multi-Indices - Combining Datasets - Merge and Join - Categories of Joins - Aggregation and Grouping - Pandas aggregation methods - Filtering - Pivot Tables - Vectorized String Operations - Methods using regular expressions –

Unit 3

Working with Time Series - Dates and Times in Python - date and time codes - Pandas Time Series: Indexing by Time - Pandas Time Series Data Structures - Frequencies and Offsets - Resampling, Shifting, and Windowing. Data Wrangling: Clean, Transform, Merge, Reshape

Text Book(s) / Reference(s)

Jake VanderPlas, Python Data Science Handbook - Essential Tools for Working with Data, O'Reilly, 2nd Edition, 2022

Wes McKiney, Python for Data Analysis, Data wrangling with Pandas, Numpy and Jupyter, O'Reilly, 3rd Edition, 2022

Paul J. Deitel, Harvey Deitel, Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and The Cloud. Pearson, 2020

Evaluation Pattern: 70:30

Assessment	Internal	End Semester
Midterm	20	-
*Continuous Assessment (Theory) (CAT)	10	-
Continuous Assessment (Lab) (CAL)	40	-
End Semester	–	30

•CA – Can be Quizzes, Assignment, Projects, and Reports

23CSE233M Database Management Systems for Data Science L-T-P-C: 2-0-3-3

Pre-Requisite(s): Nil

Course objectives

This course introduces the scope of databases and its management towards using them for data science applications. Role of SQL for data science is discussed with various real world examples. Data analysis using SQL is introduced. Techniques to deal with Text analysis and anomaly detection are illustrated. Overview of relational database on cloud is given.

Course Outcomes

After completing this course, the students will be able to

CO1: Analyze data with SQL and Python.
CO2: Analyze data for quality and take measures to deal with data quality
CO3: Apply SQL for trending, Cohort Analysis and behaviour analysis.
CO4: Apply the techniques of Text analytics using SQL.
CO5: Apply SQL for anomaly detection.

CO-PO Mapping

PO/PSO CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO1	3	3	3	2	3			1	1			2	2	1
CO2	3	3	3	2	3	1		1	1			2	1	1
CO3	3	3	3	3	3	1		1	1	2		2	2	1
CO4	3	3	3	3	3	1		1	1	2		2	2	1
CO5	3	3	3	3	3	2		1	1	2		2	2	2

Syllabus

Unit 1

Overview of Database and Database management systems – SQL – SQL for data science – Analysis with SQL – Data Analysis Workflow – Database types – Preparing data for analysis – Types of data – SQL query structure – Profiling – Distributions – Data quality – Deduplication with GROUP BY and DISTINCT - Data cleaning - Dealing with Nulls: coalesce, nullif, nvl Functions - Missing Data - Preparing: Shaping Data - BI, Visualization, Statistics, ML - Pivoting with CASE Statements - Unpivoting with UNION Statements - pivot and unpivot Functions

Unit 2

Time Series Analysis - Date, Datetime, and Time Manipulations - Trending the Data - Cohorts – Cohort Analysis – Analysis Framework - Rolling Time Windows – Sparse Data - Analyzing with

Seasonality - Retention - SQL for a Basic Retention Curve - Adjusting Time Series to Increase Retention Accuracy - Cohorts Derived from the Time Series - Defining the Cohort from a Separate Table - Dealing with Sparse Cohorts - Defining Cohorts from Dates Other Than the First Date - Related Cohort Analyses - Survivorship - Returnship, or Repeat Purchase Behavior - Cumulative Calculations - Cross-Section Analysis, Through a Cohort Lens

Unit 3

Text Analysis with SQL - What Is Text Analysis - Why SQL Is a Good Choice for Text Analysis - When SQL Is Not a Good Choice - The UFO Sightings Data Set - Text Characteristics - Text Parsing - Text Transformations - Finding Elements Within Larger Blocks of Text - Wildcard Matches: LIKE, ILIKE - Exact Matches: IN, NOT IN - Regular Expressions - Constructing and Reshaping Text - Concatenation - Reshaping Text - Database and cloud – Built-in functions – python support for accessing databases

SQL for anomaly detection - Experiment Analysis with SQL - Correlation Is Not Causation - Experiments with Binary Outcomes: The Chi-Squared Test - Experiments with Continuous Outcomes: The t-Test – Challenges - Variant Assignment – Outliers - Time Boxing - Pre-/Post-Analysis - Natural Experiment Analysis

Text Book(s) / Reference(s)

Cathy Tanimura, SQL for Data Analysis: Advanced Techniques for Transforming Data into Insights, O'Reilly Media, 2021

Richard Machina, SQL Programming For Beginners: The Guide With Step by Step Processes on Data Analysis, 2020

Upom Malik, Matt Goldwasser, Benjamin Johnston, SQL for Data Analytics: Perform fast and efficient data analysis with the power of SQL,, Packt Publishing, Year: 2019

Silberschatz. A., Korth, H. F. and Sudharshan, S., Database System Concepts‖, 6th Edition, TMH, 2010

Elmasri, R. and Navathe, S. B., Fundamentals of Database Systems‖, 5th Edition, Addison Wesley, 2006

Date, C. J. , An Introduction to Database Systems‖, 8th Edition, Addison Wesley, 2003.

Ramakrishnan, R. and Gehrke, J. ―Database Management Systems‖, 3rd Edition, McGrawHill, 2003

Evaluation Pattern: 70:30

Assessment	Internal	End Semester
Midterm	20	-
*Continuous Assessment (Theory) (CAT)	10	-
Continuous Assessment (Lab) (CAL)	40	-
End Semester	–	30

•CA – Can be Quizzes, Assignment, Projects, and Reports

23CSE234M Data Visualization L-T-P-C: 2-0-3-3

Pre-Requisite(s): Nil

Course objectives

This course highlights the importance of data visualization in data science. Overview of tools of visualization is given. Techniques to create different types of plots and charts are illustrated with real time examples. Geographic Data and its visualization are introduced. Advanced visualizations and dashboard creations are introduced using Seaborn, Tableau and Plotly.

Course Outcomes

After completing this course, the students will be able to

CO1: Use the principles of data visualization and proper tools.
CO2: Implement and display the charts and plots for real world applications using the Libraries
CO3: Create advanced visualizations using packages and tools for geometric data and Maps
CO4: Analyze real world data by designing dashboards using standard tools and packages

CO-PO Mapping

PO/PSO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO														
CO1	3	2	2	2	3	2						2	3	
CO2	3	2	2	2	3	3		2	1	1		2	3	1
CO3	3	3	3	3	3	3		2	1	2		2	3	1
CO4	3	3	3	3	3	3		2	1	2		2	3	3

Syllabus

Unit 1

Introduction to Data Visualization – Principles - Data Visualization tools - Matplotlib - How to Display the plots - Plotting from a script - Adjusting the Plot: Line Colors and Styles - Axes Limits - Labelling Plots - Simple Scatter Plots - Visualizing Errors - Density and Contour Plots - Histograms, Binnings, and Density, Kernel density estimation – Legend - Customizing Colorbars - Choosing the colormap - Sequential colormaps - Divergent colormaps - Qualitative colormaps - Color limits and extensions - Multiple Subplots - Text and Annotation - Transforms and Text Position - Arrows and Annotation - Customizing Ticks – Stylesheets – ggplot - Three-Dimensional Plotting - Contour Plots - Wireframes and Surface Plots - Surface Triangulations

Unit 2

Geographic Data with Basemap - Map Projections - Cylindrical projections - Perspective projections - Conic projections - Drawing a Map Background - Plotting Data on Maps – Visualization with Seaborn - Pair plots - Factor plots - histogram as a special case of a factor plot - violin plot

Unit 3

Basic graphs using Plotly - Tableau - Advanced visualizations with Tableau - Choropleth Maps - Waffle Charts – Dashboards – Creating Dashboards with Tableau and Plotly

Text Book(s) / Reference(s)

Jake VanderPlas, Python Data Science Handbook - Essential Tools for Working with Data, O'Reilly, 2nd Edition, 2022

Jake VanderPlas, Python Data Science Handbook - Essential Tools for Working with Data, O'Reilly , 2017

Tamara Munzner, "Visualization Analysis and Design", A K Peters Visualization Series, CRC Press, 2014

Scott Murray," Interactive Data Visualization for the Web", O'Reilly, 2013.

Alberto Cairo, "The Functional Art: An Introduction to Information Graphics and Visualization", New Riders, 2012

Nathan Yau, "Visualize This: The FlowingData Guide to Design, Visualization and Statistics", John Wiley & Sons, 2011

Evaluation Pattern: 70:30

Assessment	Internal	End Semester
Midterm	20	-
*Continuous Assessment (Theory) (CAT)	10	-
Continuous Assessment (Lab) (CAL)	40	-
End Semester	–	30

•CA – Can be Quizzes, Assignment, Projects, and Reports

23CSE235M Machine Learning with Python L-T-P-C: 2-0-3-3

Pre-Requisite(s): Nil

Course objectives

This course introduces various machine learning algorithms and their applications in data science. Tools and libraries required are introduced. Scope of using supervised and unsupervised learning techniques are introduced. Means of evaluating Machine Learning algorithms and Model Selection are introduced. Overview of Ensemble methods and Inference in Graphical Models are provided.

Course Outcomes

After completing this course, the students will be able to

CO1: Apply different types of machine learning algorithms using tools and libraries.

CO2: Apply the techniques of supervised learning and graphical models for real world applications

CO3: Apply the techniques of Pre-processing.

CO4: Apply the techniques of clustering and Dimensionality Reduction.

CO5: Analyze the performance of algorithms using different metrics and to use methods like ensemble to improve performance.

CO-PO Mapping

PO/PSO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO														
CO1	3	3	2	3	3			1	1			2	3	1
CO2	3	3	3	3	3	1		1	1	1		2	3	2
CO3	3	3	3	3	3	2		1	1	1		1	3	1
CO4	3	3	3	3	3	2		1	1	1		2	3	2
CO5	3	3	3	3	3	2		1	1	1		2	3	2

Syllabus

Unit 1

Introduction – Types of learning - Essential Libraries and Tools - scikit-learn - Jupyter Notebook - NumPy - SciPy - matplotlib - pandas – mglearn - Supervised Learning - Regression/-Classification Basic methods: Distance-based methods, Nearest- Neighbours, Decision Trees, Naive Bayes. Linear models: Linear Regression, Logistic Regression

Unit 2

Unsupervised Learning and Pre-processing - Challenges in Unsupervised Learning - Pre-processing and Scaling - Different Kinds of Pre-processing - Applying Data Transformations - Scaling Training and Test Data - The Effect of Pre-processing on Supervised Learning - Dimensionality Reduction, Feature Extraction, and Manifold Learning - Principal Component Analysis - Non-Negative Matrix Factorization - Manifold Learning with t-SNE - Clustering - k-Means Clustering - Agglomerative Clustering - DBSCAN - Comparing and Evaluating Clustering Algorithms - Generalized Linear Models.- Support Vector Machines, Nonlinearity and Kernel Methods. Beyond Binary Classification: Multi-class/Structured Outputs

Unit 3

Evaluating Machine Learning algorithms and Model Selection - Cross-Validation - Cross-Validation in scikit-learn - Benefits of Cross-Validation - Stratified k-Fold Cross-Validation and Other Strategies - Grid Search - Overfitting the Parameters and the Validation Set - Grid Search with Cross-Validation - Evaluation Metrics and Scoring - Metrics for Binary Classification - Metrics for Multiclass Classification - Regression Metrics - Using Evaluation Metrics in Model Selection

Ensemble Methods - Boosting, - Bagging - Random Forests - Reinforcement Learning, - Inference in Graphical Models - Introduction to Bayesian Learning and Inference.

Text Book(s) / Reference(s)

Jake VanderPlas, Python Data Science Handbook - Essential Tools for Working with Data, O'Reilly, 2nd Edition, 2022

Andreas C. Müller, Sarah Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, O'Reilly Media, 2016

David Julian, Designing Machine Learning Systems with Python, Packt Publishing, 2016

Bastiaan Sjardin, Luca Massaron, Alberto Boschetti, Large Scale Machine Learning with Python, Packt Publishing, 2016

Kevin Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012

Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2007.

Evaluation Pattern: 70:30

Assessment	Internal	End Semester
Midterm	20	-
*Continuous Assessment (Theory) (CAT)	10	-
Continuous Assessment (Lab) (CAL)	40	-
End Semester	–	30

•CA – Can be Quizzes, Assignment, Projects, and Reports

23CSE236M Big Data Frameworks for data science L-T-P-C: 2-0-3-3

Pre-Requisite(s): Nil

Course objectives

This course introduces the relevance of big data in data science. Frameworks to deal with the big data are introduced. Techniques and algorithms to deal with real world data are introduced. Hadoop and its applications are illustrated. Case studies and their implementations are done for streaming tweets and sentiment analysis using MongoDB, NoSQL, Spark SQL and Twitter APIs. Techniques to create a dashboard for an IoT application are introduced. Cognitive computing tools are introduced through IBM Watson.

Course Outcomes

After completing this course, the students will be able to

CO1: Use big data databases, frameworks and cloud for real time applications.

CO2: Apply MongoDB for a real world streaming data applications.

CO3: Implement and apply Hadoop for applications by creating clusters.
CO4: Apply Spark and Twitter APIs for streaming data applications
CO5: Analyze the data by creating a dashboard and to use cognitive computing resources.

CO-PO Mapping

PO/PSO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO														
CO1	2	3	2	2	3	2	1	1	2			2	1	3
CO2	2	3	3	3	3	2		1	2	1		2	1	3
CO3	2	3	3	3	3	2		1	2	1		2	2	3
CO4	2	3	3	3	3	2		1	2	1		2	2	3
CO5	3	3	3	3	3	2		1	2	1		2	3	3

Syllabus

Unit 1

Big Data: Hadoop, Spark, NoSQL and IoT - 1ntroduction - Relational Databases and Structured Query Language - NoSQL and NewSQL Big-Data Databases - NoSQL Key–Value Databases - NoSQL Document Databases - NoSQL Columnar Databases - NoSQL Graph Databases - NewSQL Databases - Case Study: A MongoDB JSON Document Database - Creating the MongoDB Atlas Cluster - Streaming Tweets into MongoDB

Hadoop - Hadoop Overview - Summarizing Word Lengths in Romeo and Juliet via MapReduce - Creating an Apache Hadoop Cluster in Microsoft Azure HDInsight - Hadoop Streaming - Implementing the Mapper - Implementing the Reducer - Preparing to Run the MapReduce Example - Running the MapReduce Job

Unit 2

Spark - Spark Overview - Docker and the Jupyter Docker Stacks - Word Count with Spark - Spark Word Count on Microsoft Azure - Spark Streaming: Counting Twitter Hashtags Using the pyspark-notebook Docker Stack - Streaming Tweets to a Socket - Summarizing Tweet Hashtags; - Introducing Spark SQL

Data Mining Twitter - Twitter APIs – Tweepy - Spotting Trends: Twitter Trends API - Twitter Streaming API - Tweet Sentiment Analysis

Unit 3

Internet of Things and Dashboards - Publish and Subscribe - Visualizing a PubNub Sample Live Stream with a Freeboard - Dashboard - Simulating an Internet-Connected Thermostat in Python - Creating the Dashboard with Freeboard.io - Creating a Python PubNub Subscriber

IBM Watson and Cognitive Computing - Introduction: IBM Watson and Cognitive Computing - IBM Cloud Account and Cloud Console - Watson Services - Additional Services and Tools - Watson Developer Cloud Python SDK - Case Study: Traveler's Companion Translation App - Test-Driving the App - SimpleLanguageTranslator.py Script - Watson Resources

Text Book(s) / Reference(s)

Scott Burk, Gary D. Miner, It's All Analytics!: The Foundations of Al, Big Data and Data Science Landscape for Professionals in Healthcare, Business, and Government, Productivity Press,2020

Paul J. Deitel, Harvey Deitel, Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and The Cloud. Pearson, 2020

Learning Spark: Lightning-Fast Big Data Analysis, Learning Spark: Lightning-Fast Big Data Analysis, O'Reilly Media, 2015

Holden Karau, Rachel Warren, High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark, O'Reilly Media, 2017

Eelco Plugge, Tim Hawkins, Peter Membrey, The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing, Apress, 2010

Shashank Tiwari, Professional NoSQL, Wrox, 2011

Evaluation Pattern: 70:30

Assessment	Internal	End Semester
Midterm	20	-
*Continuous Assessment (Theory) (CAT)	10	-
Continuous Assessment (Lab) (CAL)	40	-
End Semester	–	30

•CA – Can be Quizzes, Assignment, Projects, and Reports