**BHARATHIAR UNIVERSITY :: COIMBATORE 641 046**

**M.Sc. Data Science Curriculum (University Department)**

*(For the students admitted during the academic year 2024 – 2025 onwards)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Course Code** | **Title of the Course** | **Credits** | **Hours** | **Maximum Marks** |
| **Theory** | **Practical** | **CIA** | **ESE** | **Total** |
| **FIRST SEMESTER** |
| 24DS1C1 | Linear Algebra | 4 | 4 | - | 25 | 75 | 100 |
| 24DS1C2 | Probability and Statistics | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS1C3 | Principles of Data Science | 4 | 4 | - | 25 | 75 | 100 |
| 24DS1C4 | Database Management Systems | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS1C5 | Python and R Programming | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS1EX | Elective-I | 4 | 4 | - | 25 | 75 | 100 |
| 1GS | General Supportive -I | 2 | 2 | - | 12 | 38 | 50 |
| PDC | Industry Literacy  | 1 | - | - | 25 | - | 25 |
| 24DS1JOC1 | Job-Oriented Course  | 2 | 2 | - | 50  | - | 50 |
|  | **Total** | **29** |  |  |  |  | **725** |
| **SECOND SEMESTER** |
| 24DS2C1 | Optimization Techniques Using MATLAB | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS2C2 | Multivariate Data Analysis | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS2C3 | Machine Learning Techniques | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS2C4 | Data Mining  | 4 | 4 | - | 25 | 75 | 100 |
| 24DS2C5 | IoT Analytics | 4 | 4 | - | 25 | 75 | 100 |
| 24DS2EX | Elective-II | 4 | 4 | - | 25 | 75 | 100 |
| 2GS | General Supportive-II | 2 | 2 | - | 12 | 38 | 50 |
| 24DS2MP | Mini Project – I | 2 | - | - | 50 | - | 50 |
| 24DS2VAC1 | Value Added Course | 2 | 2 | - | 50 | - | 50 |
|  | **Total** | **30** |  |  |  |  | **750** |
| **THIRD SEMESTER** |
| **24DS3C1** | **Big Data Analytics**  | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS3C2 | Deep Learning Techniques | 4 | 2 | 4 | 25 | 75 | 100 |
| 24DS3C3 | Data Privacy and Security | 4 | 4 | - | 25 | 75 | 100 |
| 24DS3EX | Elective-III  | 4 | 4 | - | 25 | 75 | 100 |
| 24DS3EX | Elective-IV | 4 | 4 | - | 25 | 75 | 100 |
| 24DS3MP | Mini Project - II | 4 | - | - | 25 | 75 | 100 |
| 3GS | General Supportive-III | 2 | 2 | - | 12 | 38 | 50 |
| **PDC** | **Research Review Analysis** | 1 | - | - | 25 | - | 25 |
| 24DS3VAC2 | Value Added Course | 2 | - | - | 50 | - | 50 |
|  | **Total** | **29** |  |  |  |  | **725** |
| **FOURTH SEMESTER** |
| 24DS4PW | Project Work  | 12 | - | - | 180 | 120 | 300 |
| 24DS4JOC2 | Job Oriented Course | 2 | - | - | 50 | - | 50 |
|  | **Total** | **14** |  |  |  |  | **350** |
|  | **Grand Total** | **102** |  |  |  |  | **2550** |

**Note: Students must mandatorily publish or present their Research Review Analysis work in an international journal/conference before attending the viva-voce**

**Online Course**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | SWAYAM – MOOC Course\* | 2 |  |  |  |  |   |

\*Swayam – Mooc online course shall be for at least 4 weeks with at least 2 credits.

 The course shall be mandatory and shall be completed within the third semester (i.e., before the beginning of the fourth semester)

**Elective Papers**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sem** | **Elective** | **Suggested Code** | **Title of the Paper** | **No. of Credits** |
| I | Elective – I | 24DS1E1 | Design of Algorithms | 4 |
| 24DS1E2 | Artificial Intelligence | 4 |
| 24DS1E3 | Business Intelligence | 4 |
| II | Elective – II | 24DS2E1 | Transforms and Applications | 4 |
| 24DS2E2 | Predictive Analytics | 4 |
| 24DS2E3 | Software Project Management | 4 |
| III | Elective – III | 24DS3E1 | Natural Language Processing | 4 |
| 24DS3E2 | Social Media Analytics | 4 |
| 24DS3E3 | Health Care Analytics | 4 |
| 24DS3E4 | Nature Inspired Computing | 4 |
| Elective – IV | 24DS3E5 | Cloud Security | 4 |
| 24DS3E6 | Sentiment Analysis | 4 |
| 24DS3E7 | Text Analytics | 4 |
| 24DS3E8 | Digital Marketing Analytics | 4 |

**List of Job-Oriented Courses**

1. Data Analysis using Excel

2. Power BI for Data Analytics

3. Mobile Application Development

4. Smart Applications with the Internet of Things

**List of Value-Added Courses**

1. Software Testing Tools

2. Cyber Security and Digital Forensics

3. Remote Sensing and GIS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Course Code** | **24DS3C1** | **BIG DATA ANALYTICS** | **L** | **T** | **P** | **C** |
| **Core/Elective/Supportive** |  **CORE** | **2** | **0** | **4** | **4** |
| **Pre-requisite** | Basics of Big Data, Technologies, and Applications | **Syllabus Version** | **2024-25** |
| **Course Objectives:** |
| The main objectives of this course are:1. To provide in-depth knowledge about the basic concepts of Big Data, characteristics, and industry examples.
2. To discuss the Hadoop framework, HDFS, and MapReduce.
3. To inculcate HBase, Cassandra, HiveQL, Pig, and Neo4j data models.
4. To understand the need and application of MapReduce.
5. To know about the research that requires the integration of large amounts of data.
 |
|  |
| **Expected Course Outcomes:** |
| On the successful completion of the course, student will be able to: |
| 1 | Understand about basics of Big Data, Technologies and Applications in various domains. | K2 |
| 2 | Understand the foundations of Hadoop and Hadoop Distributed File System. Design of HDFS and file-based data structures. | K2/K3/ K4 |
| 3 | Analyze the working of Map Reduce and YARN for job scheduling. | K2/K3/K4 |
| 4 | Evaluate the need and fundamentals of HBase. Apply the Cassandra data model for different applications. Understand the basic commands in HiveQL, Pig and Pig Latin. | K2/K3/K4 |
| 5 | Analyze the basic concepts and need for Graph databases, create databases and retrieve records using Neo4j. Understand the data visualization and its need. | K2/K3/K4 |
| **K1** - Remember; **K2** - Understand; **K3** - Apply; **K4** - Analyze; **K5** - Evaluate; **K6** - Create |
|  |
| **Unit 1** | **Introduction to Big Data** | **10 hours** |
| Introduction: What is big data – why big data – convergence of key trends - unstructured data – industry examples of big data – Web analytics - big data and marketing – fraud and big data - risk and big data – credit risk management – big data and algorithmic trading - big data and healthcare – big data in medicine – advertising and big data – big data technologies - cloud and big data– mobile business intelligence – crowd sourcing analytics. |
|  |
| **Unit 2** | **Hadoop** | **12 hours** |
| History of Hadoop - The Hadoop Distributed File System – components of Hadoop - Analyzing the Data with Hadoop - Design of HDFS – HDFS concepts - Hadoop I/O – data integrity – compression – serialization – Avro – file-based data structures. |
|  |
| **Unit 3** | **MapReduce** | **15 hours** |
| MapReduce: MapReduce workflows – unit tests with MRUnit – test data and local tests – anatomy of MapReduce job run – classic Map-reduce – YARN – failures in classic Map-reduce and YARN – job scheduling – shuffle and sort – task execution –MapReduce types – input formats – output formats. |
|  |
| **Unit 4** | **Hadoop Eco System** | **10 hours** |
| HBase – data model and implementations – HBase clients – HBase examples. Cassandra – Cassandra data model –Cassandra examples – Cassandra clients –Hadoop integration. Pig – Grunt – pig data model – Pig Latin – developing and testing Pig Latin scripts. Hive – data types and file formats – HiveQL data definition – HiveQL data manipulation –HiveQL queries-case study. |
|  |
| **Unit 5** | **Graph Databases** | **11 hours** |
| Introduction - Neo4J - Key concept and characteristics -Modeling data for neo4j - Importing data into neo4j - visualizations - neo4j - Cypher Query Language –data visualization.  |
| **Unit 6** | **Contemporary Issues** | **2 hours** |
| Discussion on case study - Expert lectures - Online seminars – Webinars – Workshops |
|  |
|  | **Total Lecture hours** | **60 hours** |
| **Text Books** |
| 1 |  Daimi, Kevin, Hamid R. Arabnia, Principles of Data Science. Ed. Springer, 2020. |
| 2 |  SinanOzdemir, Principles of Data Science: Mathematical Techniques and Theory to Succeed in Data-Driven Industries, Packt Publishing Limited, 2016 |
| 3 | Rik Van Bruggen, Learning Neo4j, Second Edition, PacktPubishers, 2014. |
| 4 | Michael Minelli, Michelle Chambers, Ambiga Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, Wiley, 2013. |
| 5 | Tom White, “Hadoop: The Definitive Guide”, Fourth Edition, O′Reilly Publishers, 2012. |
|  |
| **Reference Books** |
| 1 | Andreas Francois Vermeulen, Ankurgupta, Cindy Gross, David Kjerrumgaard and Scott Shaw, Practical Hive: A Guide to Hadoop’s Data Warehouse System, Apress Media publishers, 2016 |
| 2 | Eric Lubow and Russell Baradberry, Practical Cassandra: A Developer’s Approach, Addison Wesley publishers, 2014. |
| 3 | EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley publishers, 2015. |
| 4 | Bart Baesens, Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, Wiley Publishers, 2015. |
| 5 | Kim H. Pries and Robert Dunnigan, Big Data Analytics: A Practical Guide for Managers, CRC Press, 2015. |
|  |
| **Related Online Contents [MOOC, SWAYAM, NPTEL, Websites etc.]** |
| 1 | TEL, Websites etc.] 1 https://nptel.ac.in/courses/106/104/106104189/ |
| 2 | https://www.edureka.co/blog/big-data-tutorial |
| 3 | https://www.coursera.org/learn/big-data-introduction |
| 4 | https://nptel.ac.in/courses/106107220 |
| Course Designed By: **Dr. D. Napoleon** |

|  |
| --- |
| **Mapping with Programme Outcomes** |
| **COs** | **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** |
| **CO1** | **S** | **M** | **S** | **S** | **M** | **L** | **M** | **L** | **M** | **L** |
| **CO2** | **S** | **M** | **M** | **S** | **S** | **M** | **S** | **L** | **M** | **M** |
| **CO3** | **S** | **L** | **S** | **M** | **S** | **L** | **L** | **M** | **S** | **M** |
| **CO4** | **S** | **M** | **S** | **M** | **M** | **M** | **S** | **S** | **M** | **M** |
| **CO5** | **S** | **L** | **S** | **M** | **M** | **S** | **M** | **S** | **S** | **M** |

**S- Strong; M-Medium; L-Low**

**List of Programs**

1. Installing Hadoop; Understanding different Hadoop modes. Startup scripts, Configuration files.
2. Hadoop Implementation of file management tasks, such as adding files and directories, retrieving files, and deleting files.
3. Data Manipulation Using HiveQL

To practice data manipulation operations in Hive using HiveQL: INSERT, UPDATE, ALTER, TRUNCATE, and DELETE.

1. Hive Join Operations

Perform the following HiveQL join operations and display the result: Inner Join, Left Outer Join, Right Outer Join, Full Outer Join

1. Implementation of Partitioning and Bucketing in Hive
2. Loading and Filtering Data in Apache Pig (load, display, join, filter, and group)
3. Aggregation and Sorting Using Apache Pig (Load, Distinct, Average (GROUP + AVG), Sort, Display)
4. Complete CRUD Operations in HBase Using HBase Shell

To perform table creation, data insertion, retrieval, updating, deletion, and scanning in HBase using HBase Shell

1. Managing HBase Table Metadata and Structure

To perform various administrative operations in HBase, including checking for table existence, listing tables, altering table schema, enabling/disabling tables, and truncating data.

1. Modeling a University Academic Network Using Neo4j

To model an academic environment using Neo4j, where people (students and professors) are linked to their departments and courses. Create and query the graph using Cypher.

1. Maintaining Product Information Using Cassandra (keyspaces and tables- INSERT, UPDATE, DELETE, and SELECT, filtering with WHERE)
2. MapReduce for Word Count using Python
3. MapReduce to Analyze Student Marks using Python

Write a MapReduce Python program to perform the following operations:

1. Calculate the average mark of each student across all subjects.
2. Identify the highest mark each student has obtained.
3. Identify the lowest mark each student has obtained.
4. Filter and list students whose average mark is above 80.