

## COURSE DELIVERY MANUAL

**(B.Tech Programme)**

|  |  |
| --- | --- |
| Name of Course with code | : CDT307: Big Data Processing |
| Offering Department | : AI & DS |
| Semester & Programme (Branch) | : Vth Sem CSE (Data Science) |
| Course Instructor , Department | : Dr. Binju Saju, AI & DS |
| Name & Designation of Stream Coordinator  Academic Year & Batch | : Dr. Binju Saju, Associate Professort    : 2025, 2023-2027 Batch |
| Type of the Course Content\* | : |
| Course project linked (Y/N) | : N |

***{\* 1.Introductory 2.Theoretical 3.Problematic 4.Analytical 5.Programming 6.Simulation 7.Designing 8.Practical/implementation.}***

## Course Overview: Big Data Processing offers a rigorous foundation in the principles, tools, and technologies essential for managing and analyzing large-scale datasets within the domain of Computer Science and Engineering (Data Science). It introduces students to the architecture and ecosystem of Hadoop, the programming capabilities of MapReduce, and the functionalities of HDFS, HBase, Pig, Hive, and Spark. Through a blend of theoretical instruction and practical implementation, learners will explore distributed storage systems, stream data models, and scalable data processing frameworks. The course emphasizes hands-on experience with R programming for data manipulation and fosters the ability to design reliable, scalable solutions for real-world big data challenges, preparing students for advanced roles in data engineering and analytics.

## Course Pre-Requisites: Database, Programming Languages

## Course Syllabus:

*(split up of lecture-tutorial-practical Hrs to be included-Sample given)*

## Course Objectives: Big Data Processing is to equip students with the conceptual understanding and practical proficiency required to manage, process, and analyze massive datasets using modern big data technologies. The course aims to develop competence in distributed computing frameworks such as Hadoop and Spark, data storage systems like HDFS and HBase, and data manipulation tools including Pig, Hive, and R. By integrating theoretical insights with hands-on experience, students will learn to design scalable solutions, implement parallel processing models, and build applications that address real-world data challenges across diverse domains.

## Course Outcomes:

After the completion of this course, students shall be able to:

|  |  |  |
| --- | --- | --- |
| **CO No.** | **Course Outcome** | **Knowledge Level** |
| CO1 | Understand big data and trivial data and build and maintain reliable, scalable, distributed systems | Understand |
| CO2 | Infer knowledge about the distributed storage and processing of   large datasets and extend the effective data storage   mechanisms using HDFS and HBase | Understand |
| CO3 | Model the distributed processing of large data sets across   clusters using simple programming models | Apply |
| CO4 | Identify the basics of stream computing and build applications using Hive | Apply |
| CO5 | Build applications using Pig and spark | Apply |

## CO-PO and CO-PSO mapping:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CO\PO &PSO** | **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PSO1** | **PSO2** | **PSO3** |
| **1** | 3 |  |  |  |  |  |  |  |  |  | 3 |  |  |  |
| **2** | 3 | 3 | 3 |  |  |  |  |  |  |  | 3 |  |  |  |
| **3** | 3 | 3 | 3 |  |  |  |  |  |  |  | 3 |  |  |  |
| **4** | 3 | 3 | 3 |  |  |  |  |  |  |  | 3 |  |  |  |
| **5** | 3 | 3 |  |  |  |  |  |  |  |  | 3 |  |  |  |

1-Slightly, 2-Moderately, 3-Strongly

## Gap in the Syllabus (Consolidated list of the gaps mentioned in the Syllabus):

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl No:** | **Description** | **Proposed Actions** | **Relevant COs, POs** |
| 1 | Minimal exposure to performance optimization and scalability challenges | Add advanced topics on Spark tuning, resource allocation, and benchmarking | CO5  PO3,  PO5,  PO11 |
| 2 | No formal training on visualization tools for big data insights | Introduce tools like Tableau, Power BI, or ggplot2 in R for data storytelling | CO2  PO10,  PO12 |
| 3 | Limited coverage of real-time data analytics and streaming frameworks | Introduce modules on Apache Kafka, Flink, or Storm for real-time processing | CO4  PO1,  PO3,  PO5 |

* 1. **Content Beyond Syllabus ( To meet Research /Industry/Professional Requirements)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl No:** | **Description** | **Proposed Actions** | **Relevant COs, POs** |
| 1 | Data Visualization and Storytelling | Introduce tools like Tableau, Power BI, and R ggplot2 for visual analytics | CO2 PO10, PO12 |

## Schedule for implementation of the course:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Topics and Subtopics | Learning Activities\*/ Learning  Assessments\*\* | CO’s | Reference No: |
|  | **Module No:1 & Contact hours required: 9** | | | |
| 1 | Introduction to Big data -Evolution of Big data, Big data characteristics | Lecture (L) | CO1 |  |
| 2 | RDBMS and Big Data | Tutorial (T) - Think   Pair Share |  |
| 3 | History of Hadoop, Hadoop Ecosystem, Core Components | Lecture (L) |  |
| 4 | Introduction to R - Features of R Language, Vectors, Filtering | Lecture (L) |  |
| 5 | Creating Matrices Applying Functions to Matrix Rows and Columns | Lecture (L) |  |
| 6 | Lists, Creating List, General List Operations | Practical session |  |
| 7 | Data Frames, Creating Data Frames | Lecture (L) |  |
| 8 | Matrix like Operations in Frames Applying Functions to Data Frames | Practical session |  |
| 9 | Reading and Writing Files. | Lecture (L) |  |
| **Module No:2 & Contact hours required: 8** | | | | |
| 1 | HDFS- Architecture | Lecture (L) | CO2 |  |
| 2 | Using HDFS Files | Tutorial (T) - Think   Pair Share |  |
| 3 | HDFS Design, Blocks, Namenodes and Data nodes | Lecture (L) |  |
| 4 | Basic File system Operations, Hadoop Specific File Types, | Lecture (L) |  |
| 5 | Anatomy of a file  read, | Lecture (L) |  |
| 6 | Anatomy of a file write. | Lecture (L) |  |
| 7 | HBase- HBase Architecture, | Lecture (L) |  |
| 8 | HBase Schema Design | Tutorial (T) - Think   Pair Share |  |
|  | | | | |
| **Module No:3 & Contact hours required: 8** | | | | |
| 1 | Data Processing with MapReduce: Execution Pipeline | Lecture (L) | CO3 |  |
| 2 | Runtime Coordination and Task Management in MapReduce, | Lecture (L) |  |
| 3 | Lecture (L) |  |
| 4 | Designing MapReduce implementations: Using MapReduce as a framework for parallel processing, | Lecture (L) |  |
| 5 | Lecture (L) |  |
| 6 | Face Recognition | Lecture (L) |  |
| 7 | Inverted Indexes Example, | Lecture (L) |  |
| 8 | Road Enrichment Example | Tutorial (T) - Think   Pair Share |  |
| **Module No:4 & Contact hours required: 8** | | | | |
| 1 | Introduction to Stream - Concepts | Lecture (L) | CO4 |  |
| 2 | Stream Data Model and Architecture, , | Lecture (L) |  |
| 3 | Sampling Data in a Stream, Filtering Streams | Lecture (L) |  |
| 4 | Counting Distinct Elements in a Stream. | Tutorial (T) - Think   Pair Share |  |
| 5 | Hive - Features, Data types and file formats, | Lecture (L) |  |
| 6 | primitive and collection data types | Lecture (L) |  |
| 7 | HiveQL, | Lecture (L) |  |
| 8 | Creating tables, Dropping Tables, Alter table. | Lecture (L) |  |
| **Module No:5 & Contact hours required: 8** | | | | |
| 1 | Pig -Installing and Execution, | Lecture (L) | CO5 |  |
| 2 | Data Model | Tutorial (T) - Think   Pair Share |  |
| 3 | Pig Latin: Structure, | Lecture (L) |  |
| 4 | Functions. | Lecture (L) |  |
| 5 | Spark- History of spark, | Tutorial (T) - Think   Pair Share |  |
| 6 | Storage layers for spark, | Lecture (L) |  |
| 7 | Core spark concepts, | Lecture (L) |  |
| 8 | RDD basics, RDD Operations. | Lecture (L) |  |

**\*Learning Activities**

## Lecture (L)

* + - 1. Black board Teaching
      2. Smart class Teaching

## Tutorial (T)

[A period of tuition or an intensive session given by a tutor or professor for one or several students, usually on a specific topic]

* + - 1. Problem solving
      2. Presentation
      3. Think Pair Share
      4. Practical session
      5. Discussion

**\*\*Learning Assessment Activities (LA) [Graded (G)/Non Graded (NG)]**

* + - * 1. Assignment
        2. Practical session
        3. Test

## Learning Assessment Activity (LA): (Sample Questions/topics can be given)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl No.** | **LA** | **Questions/topics** | **COs Mapped** | **WKs Mapped** |
| 1 | I | Assignment | CO1,CO2 | 1-4,8 |
| 2 | II | Practical session | CO1,CO4,CO5 | 1-5,8 |
| 4 | III | Test | CO1,CO2,CO3,CO4,CO5 | 1-5,8 |

## Rubrics for the Assessment of Learning Assessment Activity (LA)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Assessment Activity 1 | | Assessment Activity 2 | | Assessment Activity 3 | |
| Criteria | Weightage | Criteria | Weightage | Criteria | Weightage |
| Conceptual clarity | 5 | Hands-on proficiency with Hadoop, | 10 | Theoretical understanding, | Module Test -10CIE -25 |
| Problem-solving | 5 | Hive, Pig, Spark; | 10 | Analytical reasoning |
| Application of big data tools | 5 | execution accuracy | 10 | Application-based questions |

## CO Assessment Strategy:

|  |  |  |  |
| --- | --- | --- | --- |
| COs | Activities Planned | Type of Learning Activity | POs Mapped |
| 1 | Lecture on Hadoop Ecosystem | Conceptual understanding | PO1, PO6, PO8 |
| Assignment on R Programming |
| Internal Test Q&A |
| 2 | Lab session on HDFS and HBase | Hands-on practice and Design-based thinking | PO1, PO2, PO5 |
| Think Pair Share on schema design |
| 3 | Think Pair Share , Case study: Face Recognition | Coding and implementation,  Analytical reasoning,  Problem-solving | PO2, PO3, PO4 |
| Internal Test on parallel models |
| 4 | Stream data model simulation | Conceptual understanding | PO3, PO5, PO12 |
| HiveQL table creation exercise |
| 5 | Pig Latin scripting | Practical Based | PO3, PO5, PO11 |
| lab Spark RDD operations |

## Grading Policy: (May be Changed based on Scheme of Syllabus, Lab, Project, Seminar, etc.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Internal Assessment Marks(50)** | Internal Assessment Test 1 | 12.5 Marks | | **25**  **Marks\*** | |
| Internal Assessment Test 2 | 12.5 Marks | |
| Assignment/Module Test | **15 Marks\*\*** | | | |
|  | Assessment | Mark |  |
| Assignment | 5 |  |
| Module Test | 10 |  |
| Attendance | **10 Marks** | | | |

**\***More number of internal assessment tests can be given according to university regulation. Accordingly the marks may be taken.

\*\* Give details of 15 marks (List the LAs taken into the internal assessment and marks)

## Course Project and Assessment (if any): NIL

**Check list to be filled by Stream Coordinator**

|  |  |
| --- | --- |
|  | Yes/No |
| CDM Format is followed? |  |
| COs are properly formulated? |  |
| CO-PO Mapping formulated? |  |
| Content beyond Syllabus included? |  |
| Gaps Identification process done? |  |
| Mapping of Course Project done? |  |
| Separate CDM for add-on courses is included |  |
| Course Materials Prepared |  |
| Question Bank Prepared |  |
| MCQ Prepared |  |

|  |  |
| --- | --- |
| CO assessment Value (last academic year) (Individual CO values) |  |
| Pass percentage (last academic year) |  |

## Name and Signature of the Course Instructor: Date:

**Comments by the Stream Coordinator**

(Modifications from previous CDM, Suggestions etc...)

## Name and Signature of the Stream Coordinator: Date:

## Name and Signature of the Academic Head: Date:

**Name and Signature of the IQAC Coordinator: Date:**

**Name and Signature of the HOD: Date:**

# ANNEXURES

**Annexure I:**

**Knowledge and Attitude Profile (WK)**

**WK1:** A systematic, theory-based understanding of the natural sciences applicable to the discipline and awareness of relevant social sciences.

**WK2:** Conceptually-based mathematics, numerical analysis, data analysis, statistics and formal aspects of computer and information science to support detailed analysis and modelling applicable to the discipline.

**WK3:** A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline.

**WK4:** Engineering specialist knowledge that provides theoretical frameworks and bodies of knowledge for the accepted practice areas in the engineering discipline; much is at the forefront of the discipline.

**WK5:** Knowledge, including efficient resource use, environmental impacts, whole-life cost, reuse of resources, net zero carbon, and similar concepts, that supports engineering design and operations in a practice area.

**WK6:** Knowledge of engineering practice (technology) in the practice areas in the engineering discipline.

**WK7:** Knowledge of the role of engineering in society and identified issues in engineering practice in the discipline, such as the professional responsibility of an engineer to public safety and sustainable development.

**WK8:** Engagement with selected knowledge in the current research literature of the discipline, awareness of the power of critical thinking and creative approaches to evaluate emerging issues.

**WK9:** Ethics, inclusive behaviour and conduct. Knowledge of professional ethics, responsibilities, and norms of engineering practice. Awareness of the need for diversity by reason of ethnicity, gender, age, physical ability etc. with mutual understanding and respect, and of inclusive attitudes.

**Annexure-II**

**PROGRAM OUTCOMES (POs)**

## Engineering Graduates will be able to

**PO1: Engineering Knowledge:** Apply knowledge of mathematics, natural science, computing, engineering fundamentals and an engineering specialization as specified in **WK1 to WK4** respectively to develop to the solution of complex engineering problems.

**PO2: Problem Analysis**: Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development**. (WK1 to WK4)**

**PO3: Design/Development of Solutions:** Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required. **(WK5)**

**PO4: Conduct Investigations of Complex Problems:** Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions. **(WK8).**

**PO5: Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems. **(WK2 and WK6)**

**PO6: The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment. **(WK1, WK5, and WK7).**

**PO7: Ethics:** Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws. **(WK9)**

**PO8: Individual and Collaborative Team work:** Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.

**PO9: Communication:** Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences

**PO10: Project Management and Finance:** Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one’s own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.

**PO11: Life-Long Learning:** Recognize the need for, and have the preparation and ability for

independent and life-long learning adaptability to new and emerging technologies and

critical thinking in the broadest context of technological change. **(WK8)**

**PROGRAM SPECIFIC OUTCOMES (PSOs)**

**PSO1**: Graduates will have the ability to apply data science techniques, including data collection, cleaning, processing, analysis and visualization to extract actionable insights from large, complex datasets across various domains such as business, healthcare, and technology.

**PSO2**: Graduates will be proficient in designing, developing and deploying machine learning algorithms and predictive models, utilizing statistical, mathematical and computational techniques to solve real-world problems and enhance decision-making processes.

**CO-PO/CO-PSO JUSTIFICATION:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mapping** | **Strongly / Moderately / Slightly** | **Justification** | **WKs Mapped** | **SDGs Mapped** |
| CO1–PO1 | Strongly | Students gain foundational engineering knowledge of big data systems and architecture | WK1, WK2 | SDG 9 – Industry, Innovation |
| CO1–PO6 | Moderately | Understanding societal impact of scalable data systems supports responsible tech use | WK7 | SDG 11 – Sustainable Communities |
| CO1–PO8 | Slightly | Ethical handling of data is introduced but not deeply explored | WK7 | SDG 16 – Peace & Institutions |
| CO2–PO1 | Strongly | Deep technical understanding of HDFS and HBase enhances core engineering knowledge | WK2, WK3 | SDG 9 – Industry, Innovation |
| CO2–PO2 | Moderately | Students analyze distributed storage problems and propose schema solutions | WK4 | SDG 12 – Responsible Consumption |
| CO2–PO5 | Strongly | Use of modern tools like HBase and Hadoop aligns with tool-based learning | WK6 | SDG 9 – Industry, Innovation |
| CO3–PO2 | Strongly | MapReduce programming fosters analytical thinking and problem-solving | WK4, WK5 | SDG 4 – Quality Education |
| CO3–PO3 | Strongly | Students design parallel processing models for real-world data challenges | WK5 | SDG 9 – Industry, Innovation |
| CO3–PO4 | Moderately | Investigative thinking through case studies and algorithmic modeling | WK4 | SDG 3 – Good Health & Well-being |
| CO4–PO3 | Moderately | Stream computing and Hive-based applications involve design and implementation | WK5 | SDG 11 – Sustainable Communities |
| CO4–PO5 | Strongly | Hands-on use of Hive and stream tools enhances modern tool proficiency | WK6 | SDG 9 – Industry, Innovation |
| CO4–PO12 | Slightly | Encourages lifelong learning through evolving data technologies | WK8 | SDG 4 – Quality Education |
| CO5–PO3 | Strongly | Spark and Pig projects require solution design and optimization | WK5 | SDG 9 – Industry, Innovation |
| CO5–PO5 | Strongly | Use of Spark RDDs and Pig Latin reflects mastery of advanced tools | WK6 | SDG 7 – Clean Energy |
| CO5–PO11 | Moderately | Project-based learning introduces aspects of planning and resource management | WK8 | SDG 8 – Decent Work & Growth |