COURSE MODULE FOR THE SESSION 2025-26(EVEN SEMESTER)

Course Syllabi with CO's

Academic Year: 2025 – 2026

Department: Computer Science & Engineering - Data Science

Course Code	Course Title	Core/Elective	Prerequisite	Contact Hours			Total Hrs/ Sessions
				L	T	P	
BAD601	Big Data Analytics	Core	a foundation in statistics, mathematics programming languages like Python or R, database management, data analysis techniques, data visualization skills, and problem-solving abilities	3	-	2	40T + 20P

Objectives:

- To implement MapReduce programs for processing big data.
- To realize storage and processing of big data using MongoDB, Pig, Hive and Spark.
- To analyze big data using machine learning techniques.

Topics Covered as per Syllabus

Module -1

Classification of data, Characteristics, Evolution and definition of Big data, What is Big data, Why Big data, Traditional Business Intelligence Vs Big Data, Typical data warehouse and Hadoop environment. Big Data Analytics: What is Big data Analytics, Classification of Analytics, Importance of Big Data Analytics, Technologies used in Big data Environments, Few Top Analytical Tools , NoSQL, Hadoop.

Module -2

Introduction to Hadoop: Introducing hadoop, Why hadoop, Why not RDBMS, RDBMS Vs Hadoop, History of Hadoop, Hadoop overview, Use case of Hadoop, HDFS (Hadoop Distributed File System),Processing data with Hadoop, Managing resources and applications with Hadoop YARN(Yet Another Resource Negotiator). **Introduction to Map Reduce Programming:** Introduction, Mapper, Reducer, Combiner, Partitioner, Searching, Sorting, Compression.

Module -3

Introduction to MongoDB: What is MongoDB, Why MongoDB, Terms used in RDBMS and MongoDB, Data Types in MongoDB, MongoDB Query Language.

Module -4

Introduction to Hive: What is Hive, Hive Architecture, Hive data types, Hive file formats, Hive Query Language (HQL), RC File implementation, User Defined Function (UDF). **Introduction to Pig:** What is Pig, Anatomy of Pig, Pig on Hadoop, Pig Philosophy, Use case for Pig, Pig Latin Overview, Data types in Pig, Running Pig, Execution Modes of Pig, HDFS Commands, Relational Operators, Eval Function, Complex Data Types, Piggy Bank, User Defined Function, Pig Vs Hive.

Module - 5

Spark and Big Data Analytics: Spark, Introduction to Data Analysis with Spark. **Text, Web Content and Link Analytics**: Introduction, Text Mining, Web Mining, Web Content and Web Usage Analytics, Page Rank, Structure of Web and Analyzing a Web Graph.

TextBooks:

1. Seema Acharya and Subhashini Chellappan “Big data and Analytics” Wiley India Publishers, 2nd Edition, 2019.
2. Rajkamal and Preeti Saxena, “Big Data Analytics, Introduction to Hadoop, Spark and Machine Learning”, McGraw Hill Publication, 2019.

Reference Books

1. Adam Shook and Donald Mine, “MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems” - O'Reilly 2012
2. Tom White, “Hadoop: The Definitive Guide” 4th Edition, O’reilly Media, 2015.
3. Thomas Erl, Wajid Khattak, and Paul Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Pearson India Education Service Pvt. Ltd., 1st Edition, 2016
4. John D. Kelleher, Brian Mac Namee, Aoife D'Arcy -Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, MIT Press 2020, 2nd Edition

List of URL’s

1. https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset
2. https://www.youtube.com/watch?v=bAyrObl7TYE&list=PLEiEAq2VkUUJqp1k-g5W1mo37urJQOdCZ
3. https://www.youtube.com/watch?v=VmO0QgPCbZY&list=PLEiEAq2VkUUJqp1kg5W1mo37urJQOdCZ&in dex=4
4. https://www.youtube.com/watch?v=GG-VRm6XnNk
5. https://www.youtube.com/watch?v=JglO2Nv_92A

Course outcomes: The students should be able to:

At the end of the course, the student will be able to:

1. Identify and list various Big Data concepts, tools and applications.
2. Develop programs using HADOOP framework.
3. Make use of Hadoop Cluster to deploy Map Reduce jobs, PIG, HIVE and Spark programs.
4. Analyze the given data set and identify deep insights from the data set.
5. Demonstrate Text, Web Content and Link Analytics.

Internal Assessment Marks: 40 (3 Session Tests are conducted during the semester and Marks allotted based on average of all performances).

PRACTICAL COMPONENT OF IPCC

1. Install Hadoop and Implement the following file management tasks in Hadoop: Adding files and directories Retrieving files Deleting files and directories. Hint: A typical Hadoop workflow creates data files (such as log files) elsewhere and copies them into HDFS using one of the above command line utilities.
2. Develop a MapReduce program to implement Matrix Multiplication
3. Develop a Map Reduce program that mines weather data and displays appropriate messages indicating the weather conditions of the day.
4. Develop a MapReduce program to find the tags associated with each movie by analyzing movie lens data
5. Implement Functions: Count – Sort – Limit – Skip – Aggregate using MongoDB
6. Develop Pig Latin scripts to sort, group, join, project, and filter the data.
7. Use Hive to create, alter, and drop databases, tables, views, functions, and indexes.
8. Implement a word count program in Hadoop and Spark.
9. Use CDH (Cloudera Distribution for Hadoop) and HUE (Hadoop User Interface) to analyze data and generate reports for sample datasets

The Correlation of Course Outcomes (CO's) and Program Outcomes (PO's)

Subject Code	BAD601			Title: BIG DATA ANALYTICS								
List of Course Outcomes	PO 1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	Total
CO-1	2			2	-	-	-	-	-	-	-	-	4
CO-2	2		2	-	-	-	-	-	-	-	-	-	4
CO-3	2		2	-	-	-	-	-	-	-	-	-	4
CO-4		2		2	-	-	-	-	-	-	-	-	4
CO-5				2									2
Total	6	2	4	6									18

The Correlation of Program Specific Outcome's (PS0's) and Course Outcome (CO's)

Subject Code	BAD601		Title: BIG DATA ANALYTICS	
List of Course Outcome's	PSO1	PSO2	PSO3	Total
CO-1	2	2	-	4
CO-2	2	-	-	2
CO-3	2	-	-	2
CO-4	2		2	2
CO-5		2		
Total	-	-	-	10

Note: 3 = Strong Contribution 2 = Average Contribution 1= Weak Contribution - = No Contribution