Gyanmanjari Innovative University

Course Syllabus
Gyanmanjari Institute of Technology
Semester-7 (B.Tech)

Subject: Big Data Management – BETCE16403

Type of course: Professional Core

Prerequisite: Students should have basic knowledge of Database Management Systems, programming fundamentals, data structures, and computer networks. Basic understanding of SQL, operating system concepts, and elementary statistics will be helpful for learning Big Data Management.

Rationale:

Big Data Management is essential for understanding how large volumes of structured, semi-structured, and unstructured data are stored, processed, managed, and analyzed in modern digital systems. This subject helps students learn fundamental concepts of big data, distributed storage, processing frameworks, and data management techniques, enabling them to understand real-world data challenges and prepare for industry requirements in data-driven applications.

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks					
				Theory Marks		Practical Marks		CA	Total Marks
CI	T	P	C	ESE	MSE	V	P	ALA	
4	0	2	5	60	30	10	20	30	150

Legends: CI-Classroom Instructions; T – Tutorial; P - Practical; C – Credit; ESE - End Semester Examination; MSE- Mid Semester Examination; V – Viva; CA - Continuous Assessment; ALA- Active Learning Activities.

Course Content:

Sr. No	Course Content	Hrs.	% Weightage
1	**Introduction to Big Data Management:** Definition and concept of Big Data, Evolution of Big Data, Characteristics of Big Data, 5Vs of Big Data, Types of Big Data, Sources of Big Data, Need for Big Data Management, Traditional Data Management vs Big Data Management, Challenges in Big Data, Applications of Big Data in real-world domains.	12	20%
2	**Big Data Architecture and Storage Systems:** Introduction to Big Data architecture, Distributed computing concepts, Hadoop ecosystem overview, Hadoop Distributed File System (HDFS), HDFS architecture, NameNode and DataNode, Data replication and fault tolerance, Data lake and data warehouse concepts, Structured, semi-structured and unstructured data storage techniques.	12	20%
3	**Big Data Processing Frameworks:** Introduction to Big Data processing, Batch processing and real-time processing, MapReduce programming model, Working of MapReduce, YARN architecture, Introduction to Apache Spark, Spark architecture, RDD, DataFrame, Spark SQL, Comparison of Hadoop MapReduce and Spark, Use cases of Big Data processing frameworks.	12	20%
4	**NoSQL Databases and Big Data Querying:** Introduction to NoSQL databases, Need of NoSQL in Big Data, Types of NoSQL databases, Key-value database, Document database, Column-oriented database, Graph database, CAP theorem, Introduction to MongoDB, Cassandra and HBase, Introduction to Hive and Pig, Basic query processing in Big Data systems.	12	20%
5	**Big Data Analytics, Security and Applications:** Big Data analytics lifecycle, Data collection, Data cleaning, Data integration, Data visualization concepts, Introduction to machine learning in Big Data, Big Data security and privacy issues, Data governance, Ethical issues in Big Data, Big Data applications in healthcare, finance, education, social media, smart cities and business intelligence.	12	20%

Continuous Assessment:

Sr. No	Active Learning Activities	Marks
1	**Comparative Study of Big Data Architecture for Real-World Application:** In this task, each student has to individually select one real-world big data application domain such as healthcare analytics, smart city monitoring, banking fraud detection, e-commerce recommendation, or social media analytics. The student must analyze the expected data sources, data volume, velocity, variety, storage requirements, processing requirements, and security concerns. The student must also propose a suitable big data architecture using components such as HDFS, Spark, Hive, NoSQL database, and visualization tools. This activity must be completed individually, and the final analytical report with architecture diagram must be uploaded to the GMIU Web Portal.	10
2	**Mini Big Data Processing Workflow Design:** In this task, students have to work in a group of maximum 3 students and design a complete big data processing workflow for a selected dataset. The group must explain how the data will be collected, stored, cleaned, processed, queried, and visualized using suitable big data technologies such as Hadoop, Spark, Hive, Pig, or MongoDB. Students must also identify whether batch processing or real-time processing is more suitable for their selected problem and justify their decision. This activity must be completed as a group activity, and one combined workflow report with diagrams and justification must be uploaded to the GMIU Web Portal	10
3	**NoSQL Database Selection and Data Model Design Challenge:** In this task, each student has to individually choose a complex big data scenario such as online food delivery, hospital patient records, IoT sensor data, banking transactions, or social media network analysis. The student must compare at least two NoSQL database models such as document, column-family, key-value, and graph database, and justify the most suitable option for the selected scenario. The student must prepare a sample schema/data model, sample records, basic query requirements, and explain how scalability, availability, and consistency will be handled. This activity must be completed individually, and the database design document must be uploaded to the GMIU Web Portal.	10
	Total	30

Suggested Specification table with Marks (Theory): 60

Distribution of Theory Marks (Revised Bloom's Taxonomy)						
Level	**Remembrance (R)**	**Understanding (U)**	**Application (A)**	**Analyze (N)**	**Evaluate (E)**	**Create (C)**
Weightage %	10%	35%	25%	15%	10%	5%

Course Outcome:

After learning the course, the students should be able to:	
CO1	Understand the fundamentals and applications of Big Data Management.
CO2	Explain big data architecture, Hadoop ecosystem, and distributed storage.
CO3	Apply MapReduce, YARN, and Spark for data processing.
CO4	Analyze NoSQL databases, CAP theorem, and big data querying tools.
CO5	Evaluate analytics lifecycle, security, governance, and design big data solutions.

List of Practical

Sr. No	Description	Unit No.	Hrs.
1	Installation and setup of Google Colab-based Big Data environment. Understanding Google Drive mounting, dataset uploading, Python package installation, and basic data loading using Pandas.	01	02
2	Analyze a real-world large dataset in Google Colab and identify the 5Vs of Big Data such as Volume, Velocity, Variety, Veracity, and Value with suitable observations.	01	02
3	Perform big data file handling using Google Colab by reading and processing CSV, JSON, and text files from Google Drive and comparing structured, semi-structured, and unstructured data.	02	02
4	Install and configure PySpark in Google Colab and create a SparkSession for distributed data processing simulation.	02	02

4	Install and configure PySpark in Google Colab and create a SparkSession for distributed data processing simulation.	02	02
5	Implement basic data processing using PySpark RDD operations such as map, filter, reduce, count, collect, and word count on a text dataset.	03	02
6	Perform large dataset processing using Spark DataFrame operations including selection, filtering, grouping, sorting, aggregation, and handling missing values.	03	04
7	Execute Spark SQL queries on structured datasets in Google Colab to perform data analysis using SQL-based operations on Spark DataFrames.	03	04
8	Design and implement a NoSQL document-based data model using JSON data in Google Colab and perform insert, search, update, delete, and nested data access operations.	04	04
9	Perform big data analytics and visualization in Google Colab by cleaning a dataset, extracting insights, and representing results using suitable charts and summary tables.	05	04
10	Develop a mini big data pipeline in Google Colab by collecting a dataset, storing it in Google Drive, processing it using PySpark, performing analysis, and generating final visualization-based insights.	05	04
		Total	30

Instructional Method:

The course delivery method will depend upon the requirement of content and the needs of students. The teacher, in addition to conventional teaching methods by black board, may also use any tools such as demonstration, role play, Quiz, brainstorming, MOOCs etc.

From the content 10% topics are suggested for flipped mode instruction.

Students will use supplementary resources such as online videos, NPTEL/SWAYAM videos, e-courses, Virtual Laboratory.

The internal evaluation will be done on the basis of Active Learning Assignment.
Practical/Viva examination will be conducted at the end of semester for evaluation of performance of students in the laboratory.

Reference Books:

[1] Tom White – *Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale* – O'Reilly Media, 4th Edition, 2015.
[2] DT Editorial Services – *Big Data Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization* – Dreamtech Press, 1st Edition, 2016.
[3] Bill Chambers, Matei Zaharia – *Spark: The Definitive Guide: Big Data Processing Made Simple* O'Reilly Media, 1st Edition, 2018.
[4] Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman – *Big Data For Dummies* – Wiley, 1st Edition, 2013.
[5] Pramod J. Sadalage, Martin Fowler – *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence* – Addison-Wesley / Pearson, 1st Edition, 2012.