PARVATHANENI BRAHMAYYA SIDDHARTHA COLLEGE OF ARTS & SCIENCE
Autonomous
Siddhartha Nagar, Vijayawada-520010
Re-accredited at 'A+'by the NAAC

Offered to: M.C.A

22CA4T1: BIG DATA AND ANALYTICS

Course Descriptive and Purpose: This course is designed to assist students in comprehending the significance of big data in everyday life. It covers topics such as data storage and processing using Hadoop, gaining knowledge about contemporary database systems, utilizing Tableau for data visualization, and implementing Apache Spark through APIs, including SQL and Data Frames, for efficient data processing and analysis.

Course Objectives: The course helps the students to understand Big data and its role in Daily Life, Data Storage and Processing in Hadoop, Knowledge acquisition on Modern Databases, Visualization of Data with Tableau, Implementation of Apache Spark with API- SQL and Data Frames.

Course Outcomes:

On Successful completion the student will be able to:

CO1: What is Big Data, Big Data Analytics, MongoDB, Underneath an RDD, Changing in the realms of Big Data.

CO2: Infer about Apache Spark, Spark SQL and Data Frames, Operations, Typical Data Warehouse and Hadoop Environment.

CO3: Analyze Hive Architecture, Processing Data with Hadoop, MongoDB Query Language.

CO4: Explain Hadoop Overview, Hadoop Distributed File System, Map Reduce Programming.

CO5: Discuss Top Challenges facing Big Data, Data Types in MongoDB, Anatomy of Pig, Types of NoSQL Databases,Structuring Spark.

CO-PO MATRIX							
COURSE CODE	**CO-PO**	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**
	CO1	H		M			
	CO2	H	M				
	CO3	M	L				
	CO4	M	M	H		H	
	CO5	M		L			H

UNIT-I (12 Hours)

Types of Digital Data: Classification of Digital Data.

Introduction to Big Data: Characteristics of Data – Evolution of Big Data – Definition of Big Data – Challenges with Big Data – What is Big Data? – Other Characteristics of Data – Why Big Data? – Traditional Business Intelligence versus Big Data – Typical Data Warehouse Environment – Typical Hadoop Environment – Coexistence of Big Data and Data Warehouse – What is Changing in the realms of Big Data.

Big Data Analytics: What is Big Data Analytics – What Big Data Analytics is not? – Why this sudden Hype around Big Data Analytics? – Classification of Analytics – Greatest Challenges that Prevent Business from Capitalizing Big Data – Top Challenges facing Big Data – Why Big Data Analytics Important? – What Kind of Technologies are we looking toward to help meet the challenges posed by Big Data? – Data Science – Data Scientist – Terminologies used in Big Data Environments.

UNIT-II (12 Hours)

Hadoop: Features of Hadoop – Key advantages of Hadoop – Versions of Hadoop – Overview of Hadoop Ecosystem – Hadoop Distributions – Why Hadoop? – Why not RDBMS – RDBMS versus Hadoop – Distribution Computing Challenges – History of Hadoop – Hadoop Overview – Hadoop Distributed File System.

Processing Data with Hadoop: Managing Resource and Applications with Hadoop with YARN (Yet Another Recourse Negotiator) – Interacting with Hadoop Ecosystem.

UNIT-III (12 Hours)

Introduction to Map Reduce Programming: Introduction – Mapper – Reducer – Combiner – Partitioner – Searching – Sorting – Compression.

NoSQL: Where it is used? – What is it? – Types of NoSQL Databases – Why NoSQL? – Advantages ofNoSQL – What we miss with NoSQL? – Use of NoSQL in Industry – SQL versus NoSQL

MongoDB: What is MongoDB, Why MongoDB, Using JavaScript, Script Object Notation, Generating Unique Key, Support for Dynamic Queries, Storing Binary Data, Replication, Sharding, Updating Information in Place, Terms used in RDBMS and MongoDB, Data Types in MongoDB, MongoDB Query Language?

UNIT-IV (12 Hours

Hadoop Eco System:

Hive: What is Hive? – Hive Architecture – Hive Data Types – Hive File Format – Hive Query Language (HQL) – RC File Implementation – User Defined Function.

PIG: What is PIG? - Anatomy of Pig – Pig on Hadoop – Pig Philosophy – Use Case for Pig – Pig Latin – Data type in Pig – Running Pig – Execution Mode of Pig – HDFS Commands – Relational Operators – Eval Funtions – Complex Data Types – User Defined Functions – Parameter Substitution.

Hbase: Hbasics – Concepts – Clients – Hbase versus RDBMS.

UNIT-V (12 Hours)

Apache Spark:

Introduction to Apache Spark: A Unified Analytics – What Is Apache Spark? Unified Analytics – The Developer's Experience – Using Scala and PySpark Shell – Understanding Spark Application Concepts – Transformations – Actions and Lazy Evaluation – The Spark UI.

Apache Spark's API: What's Underneath an RDD? – Structuring Spark – The Data Frame API – The Dataset API – Data Frames Versus Datasets – When to Use RDDs – Spark SQL and the Underlying Engine.

Spark SQL and Data Frames: Introduction to built in Data Sources – Using Spark SQL in Spark Applications – SQL Tables and Views – Data Sources for Data Frames and SQL Tables : Data Frame Reader – Data Frame Writer – JSON – CSV- Images – Binary Files.

Common Data Frames and Spark SQL Operations: Unions – Joins – Windowing Spark SQL and Datasets: Working with Datasets: Creating Sample Data – Transforming Sample Data.

Prescribed Text Books			
S.No	Author	Title	Publisher
1	Seema Acharya- Subhashini Chellappan	Big Data and Analytics	Wiley Publications – Second Edition (UNIT I, II, III,IV)
2	Karau H, Konwinski A, Wendell P, Zaharia M	Learning Spark : Lightning Fast Data Analytics	O'Reilley Second Edition (UNIT V: 1 to 6 Chapters)

Reference Text Books			
S.No	Author	Title	Publisher
1	Tom White	Hadoop:TheDefinitiveGuide	O'Reilly, Yahoo Press, Third Edition
2	Bill Chambers & Matei Zaharia	SPARK:TheDefinitiveGuide	O'Reilley, 2018 Edition

3	Guller M	Big data Analytics with Spark: A Practitioner's Guide to using Spark for Large Scale Data Analysis	Apress, 2015

PARVATHANENI BRAHMAYYA SIDDHARTHA COLLEGE OF ARTS & SCIENCE
Autonomous
Siddhartha Nagar, Vijayawada-520010
Re-accredited at 'A+'by the NAAC

M.C.A **Semester:IV**

Course Code: 22CA4T1 Course Name: Big Data and Analytics

Time: 3 Hours **Max Marks: 70**

SECTION-A

Answer the following questions. (5×4=20Marks)

1. (a) Function Big Data. (CO1, L4)
   (or)
   (b) Classify the analytics. (CO1, L4)
2. (a) Compare RDBMS and Hadoop. (CO5, L2)
   (or)
   (b) Explain Key Components of Yarn? (CO4, L2)
3. (a) What is Hadoop Map Reduce? (CO5, L1)
   (or)
   (b) List the types of NoSQL Databases. (CO6, L1)
4. (a) Explain various Data Types for Hive. (CO4, L5)
   (or)
   (b) Compare HBase versus RDBMS (CO6, L5)
5. (a) What is Apache Spark? (CO2, L1)
   (or)
   (b) Define JSON. (CO2, L1)

SECTION-A

Answer the following questions. (5 × 10 = 50 Marks)

6. (a) Explain the Digital Data with examples. (CO1, L2)
   (b) Summarize the challenges faced by Bigdata. (CO6, L2)
   (or)
   (c) Explain Brewers Theorem with examples. (CO1, L2)
   (d) Explain the In-memory Analytics. (CO1, L2)
7. (a) Explain Hadoop Eco System with neat diagram. (CO5, L2)
   (or)
   (b) Explain HDFS File Systems with neat diagram. (CO5, L2)
8. (a) Make use of Map Reduce in Hadoop with example. (CO5, L3)
   (or)
   (b) Make use of File Read and File Write in Hadoop. (CO5, L3)
9. (a) Explain Hive Architecture with neat diagram. (CO4, L5)
   (or)
   (b) Explain CRUD Operations in MongoDB with examples. (CO4, L5)
   (c) Explain MongoDB import and export with examples. (CO4, L5)
10. (a) Explain TDD in Apache Spark with examples. (CO1, L5)
    (or)
    (b) Explain Common Data Frames and Distinguish between Data Frames Vs Datasets. (CO5, L5)
    (c) Explain Spark SQL Operations in spark. (CO2, L5)