22DS3E3: BIG DATA AND ANALYTICS

<table>
<tr><th>Course Name</th><td colspan="3">Big Data and Analytics</td><th>L</th><th>T</th><th>P</th><th>C</th><th>CIA</th><th>SEE</th><th>TM</th></tr>
<tr><th>Course Code</th><td colspan="3">22DS3E3</td><td>4</td><td>0</td><td>0</td><td>4</td><td>30</td><td>70</td><td>100</td></tr>
<tr><td colspan="2"><b>Year of Introduction:</b> 2021</td><td><b>Year of Offering:</b> 2022</td><td colspan="4"><b>Year of Revision</b>: No Revision</td><td colspan="4"><b>Percentage of Revision:</b> 7%</td></tr>
<tr><td colspan="11"><b>L</b>-Lecture, T-Tutorial, <b>P</b>-Practical, <b>C</b>-Credits, <b>CIA</b>-Internal Marks, <b>SEE</b>-External Marks, <b>TM</b>-Total Marks</td></tr>
</table>

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 help 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.

Specific objectives include:

1. To understand *Bigdata* and its role in *Daily Life*.
1. To know How data is *Stored* and *Processed* in Hadoop.
2. To acquire knowledge on *Modern Databases* working with MongoDB.
3. To implement Apache pig and Hive
4. To implement *Apache Spark* with *API- SQL and Data Frames*.

Course Outcomes:

CO1: This course provides a comprehensive understanding of digital data classification, introduces the concepts and evolution of Big Data, explores its characteristics and challenges, distinguishes it from traditional business intelligence, and delves into the significance of Big Data analytics and the technologies required to address its challenges, equipping students with the knowledge to navigate and harness the potential of Big Data in various domains.

CO2: Students will have a comprehensive understanding of Hadoop, including its key features, advantages, various versions, the Hadoop ecosystem, distributions, and why it is a preferred solution over RDBMS for handling big data. Students will also learn about the challenges of distributed computing, the history of Hadoop, and gain a thorough overview of the Hadoop Distributed File System. Additionally, students will acquire practical skills in managing resources and applications with Hadoop using YARN and interacting effectively within the Hadoop ecosystem.

CO3: Students will have a comprehensive understanding of MapReduce programming concepts, including Mapper, Reducer, Combiner, Partitioner, Searching, Sorting, and Compression. Additionally, students will gain insights into NoSQL databases, their applications, types, advantages, and use in industry, as well as a comparison between SQL and NoSQL, enabling them to make informed decisions in data processing and storage solutions.Working with MongoDB

CO4: Students will have a comprehensive understanding of key components within the Hadoop ecosystem, including Hive, Pig, and Hbase. They will gain insights into Hive's architecture, data types, file formats, and query language, as well as its use of RC Files and user-defined functions. Students will also learn about Pig, including its anatomy, philosophy, use cases, Pig Latin, data types, execution modes, HDFS commands, operators, functions, complex data types, and user-defined functions. Additionally, students will gain knowledge of Hbase, its concepts, clients, and how it compares to traditional RDBMS systems, enabling them to work effectively with big data technologies.

CO5: Students will be skilled in common DataFrame and Spark SQL operations like unions and joins, and they will know how to perform windowing operations. They will also be proficient in working with Datasets, creating sample data, and transforming it effectively, enabling them to work with Apache Spark for various data processing and analytics tasks.

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:The Definitive Guide	O'Reilly, Yahoo Press, Third Edition
2	Bill Chambers & Matei Zaharia	SPARK: The Definitive Guide	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

(An Autonomous College in the jurisdiction of Krishna University)

M.Sc.(Computational Data Science), First Semester

Course Name: BIG DATA AND ANALYTICS

Course Code: 22DS3E3

(w.e.f admitted batch 2022-23)

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

SECTION-A

Answer ALL questions **(5×4 = 20 Marks)**

1. (a) Function Big Data. (CO1,L4)

(or)

(b) Classify the analytics. (CO1,L4)

2. (a) Compare RDBMS and Hadoop. (CO2,L2)

(or)

(b) List the Key Components Of Yarn? (CO2,L1)

3. (a) What is Hadoop Map Reduce? (CO3, L1)

(or)

(b) List the types of NoSQL Databases. (CO3,L1)

4. (a) Explain various Data Yypes for Hive. (CO4, L2)

(or)

(b) Compare Hbase versus RDBMS (CO4,L2)

5. (a) What is Apache Spark?(CO5,L1)

(or)

(b) Define JSON. (CO5,L1)

Answer all questions.

All question carry equal marks. **5 × 10 = 50 Marks**

6. (a) Explain the Digital Data with examples. (CO1,L2) 5Marks

(b) Summarize the challenges faced by Bigdata. (CO1,L2) 5 Marks

(or)

© Explain Brewers Theorem with examples. (CO1,L2) 5 Marks

(d) Explain the In-memory Analytics. (CO1,L2) 5 Marks

7. (a) Explain Hadoop Eco System with neat diagram. (CO2,L2) 10 Marks

(or)

(b) Explain HDFS File Systems with neat diagram. (CO2 ,L2) 10 Marks

8. (a) Make use of Map Reduce in Hadoop with example. (CO3,L3) 10 Marks

(or)

(b) Make use of File Read and File Write in Hadoop. (CO3,L3) 10 Marks

9. (a) Explain Hive Architecture with neat diagram. (CO4, L2) 10 Marks

(or)

(b) Explain CRUD Operations in MongoDB with examples. (CO4,L2) 5 Marks

© Explain MongoDB import and export with examples. (CO4,L2) 5 Marks

10. (a) Explain TDD in Apache Spark with examples. (CO5,L2) 10 Marks

(or)

(b) Explain Common Data Frames and Distinguish between Data Frames Vs Datasets. (CO5, L5) 5 Marks

(b) Explain Spark SQL Operations in spark. (CO5,L5) 5 Marks