INSTITUTE OF AERONAUTICAL ENGINEERING
(Autonomous)
Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

BIG DATA MANAGEMENT								
VII Semester: CSE(DS)								
Course Code	**Category**	**Hours / Week**			**Credits**	**Maximum Marks**		
		L	**T**	**P**	**C**	**CIA**	**SEE**	**Total**
ACDD21	Core	3	0	0	3	40	60	100
Contact Classes: 48	**Tutorial Classes: Nil**	**Practical Classes: Nil**				**Total Classes: 48**		
Prerequisites: Basic knowledge of database systems, data structures, and programming using Python or Java.								

I. COURSE OVERVIEW:

This course introduces the concepts, technologies, and tools used in managing and processing big data. Students will explore data storage systems, distributed computing frameworks, and scalable database architectures that support the efficient handling of large-scale data. The course covers Hadoop, MapReduce, NoSQL databases, and modern data processing platforms like Spark. Emphasis is placed on real-time analytics, data ingestion pipelines, and handling structured, semi-structured, and unstructured data to derive actionable insights.

II. COURSE OBJECTIVES:

The students will try to learn:

I. The fundamental concepts and challenges of big data and how it differs from traditional data management.
II. The architecture and components of distributed big data processing frameworks like Hadoop and Spark.
III. Techniques for processing, storing, and querying large datasets using NoSQL and other scalable data models.
IV. The development of big data pipelines and analytics applications for real-world scenarios.

III. COURSE OUTCOMES:

CO1 **Explain** the fundamental concepts of Big Data, its lifecycle, and the limitations of traditional data management systems.
CO2 **Illustrate** the architecture and components of Hadoop and its ecosystem, including HDFS and MapReduce.
CO3 **Implement** data processing tasks using Apache Spark for batch and streaming data.
CO4 **Analyze** the design and usage of NoSQL databases such as MongoDB and Cassandra for scalable data storage.
CO5 **Apply** ETL (Extract, Transform, Load) processes using Big Data tools like Apache Sqoop, Flume, and Kafka.
CO6 **Develop** end-to-end Big Data pipelines for real-world scenarios involving ingestion, storage, and analytics.

IV. COURSE CONTENT:

MODULE I: INTRODUCTION TO BIG DATA AND ARCHITECTURE (09)
Definition of Big Data, Characteristics of Big Data – Volume, Velocity, Variety, Veracity, Big Data vs Traditional Data, Big Data Lifecycle, Big Data Applications, Architecture of Big Data Systems, Challenges in Big Data Management

MODULE II: HADOOP ECOSYSTEM AND HDFS (10)
Hadoop Overview, HDFS Architecture and Features, Data Flow in HDFS, MapReduce Programming Model, YARN Architecture, Hadoop Ecosystem Components – Pig, Hive, HBase, ZooKeeper, Case Study

MODULE III: SPARK AND REAL-TIME PROCESSING (10)
Limitations of MapReduce, Apache Spark Architecture, RDDs and Data Frames, Spark SQL.

Spark Streaming Basics, Structured Streaming, Comparison of Batch and Streaming Data Processing, Hands-on with Spark APIs

MODULE IV: NOSQL DATABASES (09)
Introduction to NoSQL, Types of NoSQL Databases – Key-Value, Column-Family, Document, Graph, MongoDB: CRUD Operations, Indexing, Aggregation, Cassandra Data Model and Architecture, CAP Theorem

MODULE V: DATA INGESTION, INTEGRATION, AND ANALYTICS (10)
Data Ingestion Tools – Sqoop, Flume, Kafka, ETL Processes in Big Data, Data Warehousing vs Data Lakes, Integration of BI with Big Data, Use of Data Visualization Tools, Case Study: Big Data Pipeline Implementation

V. TEXTBOOKS:

1. Viktor Mayer-Schönberger & Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013
2. Tom White, Hadoop: The Definitive Guide, 4th Edition, O'Reilly Media, 2015
3. Bill Chambers and Matei Zaharia, Spark: The Definitive Guide, O'Reilly Media, 2018
4. Shashank Tiwari, Professional NoSQL, Wiley, 2011

VI. REFERENCE BOOKS:

1. **Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman**, *Mining of Massive Datasets*, Cambridge University Press, **2014**
2. **Chuck Lam**, *Hadoop in Action*, Manning Publications, **2010**

VII. WEB REFERENCES

1. NPTEL – Big Data Computing – IIT Patna
2. Apache Hadoop Official Documentation
3. Apache Spark Documentation
4. MongoDB University – Free NoSQL and MongoDB Courses

VIII. MATERIALS ONLINE

1. Course template
2. Tutorial question bank
3. Tech talk topics
4. Open-ended experiments

5. Definitions and terminology
6. Assignments
7. Model question paper – I
8. Model question paper – II
9. Lecture notes
10. PowerPoint presentation
11. E-Learning Readiness Videos (ELRV)