III B.Tech II Semester

23A39603	BIG DATA (Professional Elective-II)	L	T	P	C
		3	0	0	3

Course Objectives:

1. To introduce the fundamental concepts and characteristics of Big Data.
2. To equip students with knowledge of Hadoop, Spark, and distributed computing systems.
3. To develop the ability to process, store, and analyze Big Data using modern tools.
4. To implement various big data analytics techniques on real-world data.
5. To apply big data tools and frameworks in domains such as finance, health, and e-commerce.

Course Outcomes:
Upon completion of the course, students will be able to:

1. Understand the characteristics, challenges, and tools of Big Data.
2. Implement storage and retrieval mechanisms in HDFS and NoSQL databases.
3. Develop MapReduce and Spark-based applications.
4. Apply big data analytics techniques to process real-world large-scale datasets.
5. Integrate big data platforms with machine learning and business intelligence solutions.

UNIT I: Introduction to Big Data and Hadoop Ecosystem
Definition, Characteristics of Big Data (Volume, Variety, Velocity, Veracity, Value), Types of Data: Structured, Semi-Structured, and Unstructured, Traditional vs Big Data Systems, Big Data Challenges and Benefits, Introduction to Hadoop: Architecture and Components, Hadoop Distributed File System (HDFS): Features, Design, Blocks, YARN and MapReduce Overview, Hadoop Ecosystem Components: Pig, Hive, HBase, Sqoop, Flume.

UNIT II: MapReduce Programming and Hadoop Tools
MapReduce Programming Model: Mapper, Reducer, Partitioner, InputSplit and RecordReader, Combiner, Writing MapReduce Programs in Java, Advanced MapReduce Concepts: Counters, Joins, Secondary Sort, Hive: Data Warehousing Concepts, HiveQL, Partitions, Buckets, Pig: Data Flow, Pig Latin Scripts, Data Import & Export with Sqoop, Real-Time Data Collection using Flume.

UNIT III: NoSQL Databases and HBase
Introduction to NoSQL Databases, Types of NoSQL: Key-Value, Document, Column, Graph, Differences between RDBMS and NoSQL, HBase Data Model: Column Families, Regions, Tables, HBase Architecture and Internals, HBase CRUD Operations using Java, Integration of HBase with Hadoop, Case Study: Big Data Storage in Social Media.

UNIT IV: Apache Spark and Big Data Analytics
Apache Spark: RDDs and DAG Execution Model, Spark Core and Spark SQL, DataFrames and Datasets in Spark, Spark Streaming: Architecture and DStreams, Spark MLlib: Machine Learning on Big Data, GraphX: Graph Processing in Spark, Performance Tuning and Optimization in Spark, Case Study: Building a Spark Application for Real-Time Analytics.

UNIT V: Applications and Case Studies in Big Data
Big Data in Healthcare: Predictive Analysis, Genomics, Big Data in Finance: Fraud Detection, Risk Analytics, Big Data in E-Commerce: Customer Behavior, Personalization, Sentiment Analysis using Big Data, Big Data for Smart Cities and IoT, Big Data and Cloud Computing Integration (AWS, GCP, Azure), Data Privacy, Security, and Ethical Issues – (E), Mini-Project: Design and Development of a Big Data Solution.

Textbooks:

1. Tom White, Hadoop: The Definitive Guide, O'Reilly Media.
2. V. Srinivasa Subramanian, Big Data Analytics, Wiley India.
3. Anand Rajaraman and Jeffrey D. Ullman, Mining of Massive Datasets, Cambridge University Press.

Reference Books:

1. Chuck Lam, Hadoop in Action, Manning Publications.
2. Bill Franks, Taming the Big Data Tidal Wave, Wiley.
3. Alex Holmes, Hadoop in Practice, Manning Publications.
4. Michael Minelli, Big Data, Big Analytics: Emerging Business Intelligence, Wiley.

Online Learning Resources:

- Coursera – Big Data Specialization by UC San Diego
- edX – Big Data Analysis with Apache Spark (UC Berkeley)
- Udacity – Data Engineering Nanodegree