# BIG DATA TECHNOLOGIES

## III B. TECH- I SEMESTER:

| Course Code: | Category | Hours / Week | | | Credits | Maximum Marks | | |
|---|---|---|---|---|---|---|---|---|
| | | L | T | P | C | CIA | SEE | Total |
| A6DS06 | PCC | 3 | 0 | 0 | 3 | 40 | 60 | 100 |
| Contact Classes:61 | Tutorial Classes: | Practical Classes: | | | | Total Classes:61 | | |

### COURSE OBJECTIVES

1. Understand what big data is and how Big Data Technologies can help organizations achieve a competitive advantage.
2. Provide an overview of Apache Hadoop and its ecosystem components.
3. Understand Map Reduce Jobs
4. Processing of Big Data with advanced architectures like Spark.
5. To understand practical machine learning scalable and easy.

### COURSE OUTCOMES

**At the end of the course, student will be able to:**

1. Understand fundamentals of Big Data Technologies.
2. Investigate Hadoop framework and Hadoop Distributed File system.
3. Demonstrate the MapReduce programming model to process the big data along with Hadoop tools.
4. Implement Big Data code in Apache Spark (in PySpark).
5. Run Supervised and Unsupervised machine learning on Large-Scale Data.

| UNIT-I | Introduction to Big Data Analytics | | Classes :12 |
|---|---|---|---|

**Introduction to Big Data Analytics:** Big Data, Scalability and Parallel Processing, Designing Data Architecture, Data Sources, Quality, Pre-Processing and Storing, Data Storage and Analysis, Big Data Analytics Applications and Case Studies.

| UNIT-II | Introduction to Hadoop | Classes-12 |
|---|---|---|

**Introduction to Hadoop:** Introduction, Hadoop and its Ecosystem, Hadoop Distributed File System, MapReduce Framework and Programming Model, Hadoop Yarn, Hadoop Ecosystem Tools.

**Hadoop Distributed File System Basics:** HDFS Design Features, Components, HDFS User Commands.
**Hadoop Ecosystem Components:** Using Apache Pig, Hive, Sqoop, Flume, Oozie, HBase.

| UNIT-III | MapReduce, Hive and Pig | Classes: 12 |
|---|---|---|

**MapReduce, Hive and Pig:** Introduction, MapReduce Map Tasks, Reduce Tasks and MapReduce Execution, Composing MapReduce for Calculations and Algorithms, Hive, HiveQL, Pig.

| UNIT-IV | Large-Scale Data Processing with PySpark | Classes:12 |
|---|---|---|

Apache Spark, Spark programming. (Python and PySpark), RDDs, Data Frames, Spark SQL, PySpark, NumPy, SciPy, Code Optimization, Cluster Configurations, Linear Algebra Computation in Large Scale, Distributed File Storage Systems.

| UNIT-V | Large Scale Machine Learning with Spark | Classes:13 |
|---|---|---|

Basic statistics, Data sources, Pipelines, Extracting, transforming and selecting features, Classification and Regression, Clustering, Collaborative filtering, Frequent Pattern Mining, Model selection and tuning.

**Text Books:**

1. Raj Kama! and Preeti Saxena, "Big Data Analytics Introduction to Hadoop, Spark, and Machine-Learning", McGraw Hill Education, 2018 ISBN: 9789353164966, 9353164966
2. Douglas Eadline, "Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem", 1st Edition, Pearson Education, 2016. ISBN13: 978-9332570351

**Reference Books:**

1. Tom White, "Hadoop: The Definitive Guide", 4th Edition, O"Reilly Media, 2015.ISBN-13: 978-9352130672
2. Perrin, J. (2020). Spark in action (2nd ed.). (Covers Apache Spark 3 with examples in Java, Python, and Scala) O'Reilly Media Inc.
3. Arshdeep Bahga, Vijay Madisetti, "Big Data Analytics: A Hands-On Approach", 1st Edition, VPT Publications, 2018. ISBN-13: 978-0996025577
4. Damji, J., Wenig, B., Das, T., Lee, D. (2020). Learning spark (2nd ed.) O'Reilly Media Inc.
5. Nudurupati, S. (2021). Essential PySpark for scalable data analytics: A beginner's guide to harnessing the power and ease of PySpark 3 Packt Publishing

**Web references:**