

PRASAD V. POTLURI SIDDHARTHA INSTITUTE OF TECHNOLOGY

(Autonomous) Kanuru, Vijayawada-520007

Department of Computer Science & Engineering (Data Science)

IV B. Tech – I Semester

BIG DATA ANALYTICS

Course Code	23DS3701	Year	IV	Semester	I
Course Category	PC	Branch	CSE (DS)	Course Type	Theory
Credits	3	L-T-P	3-0-0	Pre requisites	Database Management Systems
Continuous Internal Evaluation	30	Semester End Examination	70	Total Marks	100

Course Outcomes

Upon successful completion of the course, the student will be able to

CO1	Describe the characteristics, challenges and lifecycle of big data to understand their role in large-scale data processing.	L2
CO2	Use the knowledge of Hadoop tools (MapReduce, Pig and Hive) to process and analyze large datasets.	L3
CO3	Apply the Spark's processing concepts to perform scalable and real-time data analytics.	L3
CO4	Analyze the features, advantages, and architectures of different big Data processing frameworks to extract meaningful insights for data-driven decision making.	L4

Contribution of Course Outcomes towards achievement of Program Outcomes & Strength of correlations (**3: Substantial, 2: Moderate, 1: Slight**)

[illegible]

Syllabus		
UnitNo.	Contents	Mapped CO
I	Big Data: Definition, Characteristics of Big Data, Challenges of Big Data, Data Analytics Life Cycle, Classification of Data Analytics. Working with Big Data: History of Hadoop, HDFS, Components of HDFS, Building blocks of Hadoop, Processing the Data With Hadoop.	CO1
II	Map Reduce: Understanding Hadoop Functions, Processing Data with Map Reduce, Types of Maps Reduce, Uses of Map Reduce. Map Reduce Algorithms: Basic programs of Hadoop Map Reduce, Matrix Vector Multiplication, Matrix Multiplication, Map Tasks, Reduce tasks, Map Reduce Example- Word Count, Page Rank.	CO1, CO2, CO4
III	Hadoop Tools-PIG: Introduction to PIG, Hadoop Programming Made Easier Admiring the Pig Architecture, PIG Data models, Count, PIG Latin, Going with the Pig Latin Application Flow. Hadoop Data with Hive: Getting Started with Apache Hive, Hive Shell, Hive Services, Hive QL, Hive DDL Tables, User Defined Functions, Difference Between PIG & Hive.	CO1, CO2, CO4
IV	Introduction to Spark: Introduction to Spark Concept, Key features, advantages, Spark Ecosystem, Spark Architecture and components, Cluster Management, Comparison with Hadoop and MapReduce. Spark Resilient Distributed Datasets (RDDs): Introduction to RDDs and their characteristics, Creating RDDs, Transformations, Actions, Persistence and Caching, RDD Partitioning and Shuffling.	CO1, CO3, CO4
V	Spark Data Frames and Datasets: Introduction to Data Frames and Data sets, Differences between RDDs, Data Frames, Datasets, Creating and transforming Data Frames, Performing operations on Data Frames. Introduction to Spark SQL: Role of Spark SQL in the Spark ecosystem, SQL Context, Hive Context, Writing SQL queries to interact with Data Frames, Data Sources and File Formats.	CO1, CO3, CO4
Learning Resources		
Text Books		
1. Hadoop: The Definitive Guide, Tom White, 4th Edition, 2015, O'Reilly Media Inc. 2. Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, Bart Baesens, 1st Edition, 2014, John Wiley & Sons, Inc. 3. Learning Spark: Lightning-Fast Big Data Analytics, Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, 1st Edition, 2015, O'Reilly Media Inc. 4. Big Data Analytics: Introduction to Hadoop, Spark, and Machine Learning, Raj Kamal & Preeti Saxena — McGraw Hill Education 5. Spark: The Definitive Guide, Bill Chambers & Matei Zaharia — O'Reilly Media 6. Seema Acharya & Subhashini Chellappan – Big Data Analytics, Wiley India, 2015		
References		
1. Mining of Massive Data Sets, <u>Jure Leskovec</u>, Anand Rajaram, David Ullman, 2nd Edition 2016, Dreamtech Press. 2. Big Data Analytics with R and Hadoop, Vignesh Prajapati, I edition, 2014, Shroff Publishers & Distributors Pvt Ltd. 3. Spark: The Definitive Guide, Bill Chambers and Matei Zaharia, First Edition, 2018, O'ReillyMedia. 4. Programming Pig (2nd Edition), Alan Gates & Daniel Dai — O'Reilly Media 5. High Performance Spark, Holden Karau & Rachel Warren — O'Reilly Media		
e-Resources and other Digital Material		
1. Apache Spark Documentation & Tutorials : https://spark.apache.org/ 2. https://www.udemy.com/courses/search/?src=ukw&q=Apache+Spark+%26+Hadoop 3. https://hadoop.apache.org/ 4. https://cwiki.apache.org/confluence/display/Hive		

