

PRASAD V. POTLURI SIDDHARTHA INSTITUTE OF TECHNOLOGY

(Autonomous)

Kanuru, Vijayawada-520007

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)**III B.Tech – II Semester CSE (Data Science)****Data Analytics Lab**

Course Code	20DS3651	Year	III	Semester	II
Course Category	PCC Lab	Branch	CSE (Data Science)	Course Type	Practical
Credits	1.5	L-T-P	0-0-3	Pre requisites	Java Programming
Continuous Internal Evaluation	15	Semester End Examination	35	Total Marks	50

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

CO1	Demonstrate experimental procedures through oral communication and submit Comprehensive documentation reports.	L2
CO2	Implement data processing and analysis tasks using various components of the Hadoop ecosystem, including HDFS, MapReduce, Apache Pig, and Apache Hive.	L3
CO3	Analyze the performance and scalability of different big data processing techniques and frameworks, identifying their strengths and weaknesses in handling Various types of data and workloads.	L4
CO4	Evaluate the accuracy, efficiency, and reliability of data analytics solutions by Testing and validating them against real-world data sets and benchmarks.	L5

Contribution of Course Outcomes towards achievement of Program Outcomes & Strength of correlations(3: High,2: Medium, 1: Low)

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO1	2									2				
CO2					3							3	2	
CO3		3										3		
CO4				3								3		

Syllabus

Expt No	Contents	Mapped CO
1.	Exploring of different Big Data Analytics tools(Apache Hadoop, Apache Spark, Apache Kafka, etc)	CO1
2.	Big Data and Hadoop Ecosystem <ul style="list-style-type: none"> • Setup a single-node Hadoop cluster on a virtual machine or cloud environment • Perform basic operations on the Hadoop Distributed File System(HDFS) 	CO1 to CO4
3.	Map Reduce Programming <ul style="list-style-type: none"> • Implement the Word Count example using the Map Reduce framework • Develop a MapReduce program to find the maximum and minimum values in a dataset 	CO1 to CO4
4.	Map Reduce Programming <ul style="list-style-type: none"> • Implement Page Rank Algorithm using the Map Reduce Framework 	CO1 to CO4
5.	Apache Pig <ul style="list-style-type: none"> • Learn Pig Latin scripting language • Use Pig to load data, perform filtering, and generate summary statistics • Implement a Pig script to join and process multiple datasets 	CO1 to CO4
6.	Apache Hive <ul style="list-style-type: none"> • Create Hive tables and load data. • Write Hive QL queries for data exploration and analysis • Use Hive's built-in functions and user-defined functions 	CO1 to CO4
7.	Spark RDDs <ul style="list-style-type: none"> • Create RDDs from various data sources (files, collections, etc.) • Perform transformations and actions on RDDs • Implement caching and persisting RDDs for efficient computations 	CO1 to CO4
8.	Spark Data Frames and Datasets <ul style="list-style-type: none"> • Create Data Frames from various data sources • Perform data manipulation and transformation using Data Frames • Convert RDDs to Data Frames and vice versa 	CO1 to CO4
9	Spark SQL <ul style="list-style-type: none"> • Create temporary views and query Data Frames using Spark SQL • Integrate Spark SQL with Hive Meta store • Perform advanced data analysis using Spark SQL functions 	CO1 to CO4

10	CapstoneProject-1:(Facebook/X/etc) <ul style="list-style-type: none"> • Identify a real-world dataset • Perform data ingestion, Pre processing, and Analysis using Hadoop and Spark • Develop a end-to-end big data analytics pipeline, Integrating multiple components 	CO1 to CO4
11	CapstoneProject-2:(RecommenderSystems:Netflix/Douban/MovieLens/etc) <ul style="list-style-type: none"> • Identify a real-world dataset • Perform data ingestion, Preprocessing, and Analysis using Hadoop and Spark • Develop a end-to-end big data analytics pipeline, Integrating multiple components 	CO1 to CO4
12	CapstoneProject-3:(Banking/stock market/Insurance/etc) <ul style="list-style-type: none"> • Identify a real-world dataset • Perform data ingestion, Preprocessing, and Analysis using Hadoop and Spark • Develop a end-to-end big data analytics pipeline, Integrating multiple components 	CO1 to CO4

Learning Resources

Text Books

1. Hadoop Map Reduce Cook book, Srinath Perera &Thilina Gunarathne, 2013,PACKTPUBLISHING.
2. Big and Hadoop Learn by examples by Mayank Bhushan, FirstEdition,2018,BPB Publications.
3. Big Data Analytics with Spark: A Practitioner's Guide, Mohammed Guller ,2021,APress.
4. Hadoop: The Definitive Guide, Tom White, 2015, O'Reilly Media.
5. Learning Spark: Lightning-Fast Big Data Analytics, Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia, 2015, O'Reilly Media.

e-Resources & other digital material

1. <https://www.youtube.com/watch?v=G0xyw1ODi5A>
2. <https://www.youtube.com/watch?v=q8INOCrU9HE>
3. https://www.youtube.com/watch?v=2N9gP119_F4