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)

Big Data Analytics

Course Code	20DS3602	**Year**	III	**Semester**	II
Course Category	PCC	**Branch**	CSE(Data Science)	**Course Type**	Theory
Credits	3	**L-T-P**	3-0-0	**Pre requisites**	Data Mining, Data Bases
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	**L2**
CO2	Apply the knowledge of Hadoop tools such as MapReduce, Pig and Hive to process and analyze large datasets.	**L3**
CO3	Apply the Spark's capabilities for efficient data processing and analysis.	**L3**
CO4	Analyze and compare the features, advantages, and architectures of different big Data processing frameworks to select the most appropriate solution for given data analytics problems or use cases.	**L4**

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													
CO2	3											1	3	
CO3	2											1		
CO4		3										1		

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. **Hadoop:** History of Hadoop, HDFS, Components of HDFS, Processing the Data With Hadoop.	**CO1**
II	**Map Reduce:** Understanding Map Reduce Functions, Processing Data with Map Reduce, Map Reduce Example- Word Count, Page Rank, Types of Maps Reduce, Uses of MapReduce, MapReduce Algorithms - MatrixVector Multiplication, Matrix Multiplication, Map Tasks, Reduce tasks.	**CO1, CO2, CO4**
III	**Hadoop Tools:** Introduction to PIG, PIG Data models, Count, PIG Latin, Hive, Hive Shell, Hive Services, Hive QL, Hive DDL Tables, User Defined Functions, Difference Between PIG & Hive.	**CO1, CO2, CO4**
IV	**Introduction to Spark:** Overview, Key features, advantages, Spark Ecosystem, Spark Architecture, Cluster Management, Comparison with Hadoop and MapReduce. **Resilient Distributed Datasets(RDDs)**:Introduction to RDDs and their characteristics, Creating RDDs, Transformations, Actions, Persistence and Caching, RDD Partitioning and Shuffling.	**CO1, CO2, CO3**
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, SQLContext, HiveContext, Writing SQL queries to interact with Data Frames, Data Sources and FileFormats.	**CO1, CO3, CO4**

LearningResources

Text Books

1. Hadoop: The Definitive Guide, Tom White, Fourth Edition, 2015,O'Reilly Media Inc.
2. Analytics in Big Data World: The Essential Guide to Data Science and its Applications, Bart Baesens, 2014, Wiley.
3. Learning Spark: Lightning-Fast Big Data Analytics, Holden Karau, Andy Konwinski, Patrick Wendell and Matei Zaharia, 2015, O'Reilly Media.

References

1. Mining of Massive Data Sets, Anand Rajaram anand David Ullman,CambridgeUniversity Press, 2014.
2. Big Data Analytics with R and Hadoop, Vignesh Prajapati, 2013,SPD.
3. Spark: The Definitive Guide, Bill Chambers and Matei Zaharia, 2018, O'ReillyMedia.

e-Resources and other Digital Material

1. https://bigdatauniversity.com/courses/spark-overview-scala-analytics/
2. https://bigdatauniversity.com/courses/introduction-to-hadoop/
3. https://bigdatauniversity.com/courses/mapreduce-and-yarn/