PRASAD V. POTLURI SIDDHARTHA INSTITUTE OF TECHNOLOGY
(Autonomous) Kanuru, Vijayawada-520007
IV B. Tech – Semester-1

BIG DATA ANALYTICS

Course Code	23IT4702B	**Year**	IV	**Semester**	I
Course Category	Professional Elective-V	**Branch**	IT	**Course Type**	Theory
Credits	3	**L – T – P**	3-0-0	**Prerequisites**	Database
Continuous Evaluation	30	**Semester End Evaluation**	70	**Total Marks**	100

Course Outcomes		
Upon successful completion of the course, the student will be able to:		
CO1	Describe the significance and fundamental concepts of big data to understand their role in large-scale data processing.	L2
CO2	Apply NoSQL distributed storage techniques and Apache Spark processing concepts to perform scalable and real-time Big Data analytics.	L3
CO3	Apply the knowledge of Hadoop tools (Map Reduce, Hive) to process and analyze large datasets.	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)

	PO1	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO10**	**PO11**	**PSO1**	**PSO2**
CO1	3											2	2
CO2	3	2	2	2	1							2	2
CO3	3	2	2	2	1							2	2
CO4	3	2	2	2	1							2	2

Syllabus		
Unit No.	**CONTENTS**	**Mapped CO**
I	**What Is Big Data and Why Is It Important**? – Introduction to Big data, A Convergence of Key Trends, Unstructured Data. **Industry Examples of Big Data**- Web analytics, Big data and Marketing, Fraud and Big data, Risk and Big data, Credit Risk management, Big Data and Algorithmic Trading, Big Data and Health care, Medicine, Three Big Data Vs in Advertising. **Big Data Technology**- Introduction to Hadoop, Open-Source Technology for Big Data Analytics, The Cloud and Big Data, Mobile Business Intelligence, Crowd Sourcing Analytics, Inter- and Trans-Firewall Analytics.	**CO1**
	Introduction to NoSQL- Aggregate data models- Aggregates, key-value and document data models. **More Details on Data Models**- Relationships, graph databases, schema less	**CO1, CO2, CO4**

II	databases, materialized views. **Distribution Models**- Sharding, Master-slave replication, Peer- Peer replication, Combining sharding and replication. **Consistency**- Relaxing consistency. Version stamps.	
III	**Map Reduce-** A Weather Dataset- Data format, Analyzing data with Hadoop, Scaling Out. **Hadoop Distributed File System**-The Design of HDFS, HDFS Concepts: Blocks, Name nodes and Data nodes HDFS Federation, HDFS High - Availability. The Java Interface, Data Flow. (Text Book- 3) **Introduction to Hive-** Data Types and File formats, HiveQL: Data Definition, HiveQL: Data manipulation, HiveQL: Queries- Join Statements. Table partitioning, Bucketing.	**CO1, CO3, CO4**
IV	**A Gentle Introduction to Spark**- Spark's Basic Architecture, Spark's Language APIs, The Spark Session, Data Frames, Transformations- Lazy Evaluation. **Working with Different Types of Data**- Working with Dates and Timestamps, working with Nulls in Data. Joins. **Resilient Distributed Datasets (RDDs)-** About RDDs, Creating RDDs, Manipulating RDDs, Transformations, Actions. **Distributed Shared Variables**- Broadcast Variables, Accumulators.	**CO1, CO2, CO4**
V	**Spark**: **Streaming-** Stream Processing Fundamentals, Structured Streaming Basics - Core Concepts, Structured Streaming in Action, Transformations on Streams, Input and output. **Event-Time and State Full Processing** - Event Time, State full Processing	**CO1, CO2, CO4**

Learning Resources

Text Books

1. Big Data, Big Analytics: Emerging, Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, 1st edition ,2013.
2. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World Polyglot Persistence", Addison-Wesley Professional, 2012.
3. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012
4. "Programming Hive", O'Reilley, E. Capriolo, D. Wampler, and J. Rutherglen, 2012.
5. SPARK: The Definitive Guide, Bill Chambers &Matei Zaharia, O' Reilley, 2018-first Edition.
6. Business Intelligence and Analytic Trends for Today's Businesses", Wiley, First edition-2013.
7. Big Data Analytics: Introduction to Hadoop, Spark, and Machine Learning, Raj Kamal & Preeti Saxena — McGraw Hill Education

Reference Books

1. "Hadoop Operations", O'Reilley, Eric Sammer, First Edition -2012.
2. Mining of Massive Data Sets, Jure Leskovec, Annand Raja ram, David Ullman, 2nd Edition 2016, Dreamtech Press.
3. Big Data Analytics with R and Hadoop, Vignesh Prajapati, I edition,2014, Shroff Publishers & Distributors Pvt Ltd
4. "HBase: The Definitive Guide", O'Reilley, Lars George, September 2011: First Edition
5. "Programming Pig", O'Reilley, Alan Gates, October 2011: First Edition

E-Resources & other digital material
1. https://nptel.ac.in/courses/106104189 2. https://www.coursera.org/specializations/big-data 3. https://www.edx.org/course/big-data-fundamentals 4. https://www.coursera.org/learn/big-data-analytics-1 5. https://www.udemy.com/topic/big-data/?srsltid=AfmBOooExG-T-NX2t7zpQ3BCrPN_hH8K6MeLjncV3dJrd8w3WRIn9xIe