SARVAJANIK UNIVERSITY
Sarvajanik College of Engineering and Technology
Bachelor of Technology

Year: B. Tech III (Semester VI)

Subject Name: Big Data Analytics **Subject Code:** BTAI14601
Type of course: Professional Elective Course
Prerequisite (if any): Database Management Systems, Programming for Problem Solving

Rationale: Today's data-driven world increasingly, the efficient operation of organizations across sectors relies on the effective use of vast amounts of data. Big data analytics helps us to manage and examine these data with hidden patterns, correlations, and other insights. To understand the key issues in big data analytics and its associated applications in business analytics. Ability to understand different framework of big data.

Teaching and Examination Scheme:

Teaching Scheme				Theory Marks			Practical Marks		Total
L	T	P	C	TEE	CA1	CA2	TEP	CA3	
3	0	2	4	60	25	15	30	20	150

CA1: Continuous Assessment (assignments / projects / open book tests / closed book tests) CA2: Sincerity in attending classes / class tests / timely submissions of assignments / self-learning attitude / solving advanced problems TEE: Term End Examination TEP: Term End Practical Exam (Performance and viva on practical skills learned in course) CA3: Regular submission of Lab work / Quality of work submitted / Active participation in lab sessions / viva on practical skills learned in course.

Contents:

Sr. No.	Contents	Total Hrs
1	**Introduction to Big Data** Introduction to Big Data, Big Data characteristics, Challenges of Conventional System, Types of Big Data, Intelligent data analysis, Traditional vs. Big Data business approach, Challenges in Big Data Analytics – Need of big data frameworks	04
2	**Hadoop Framework** **Hadoop**: Requirement of Hadoop Framework - Design principle of Hadoop –Comparison with other system - Components of Hadoop Analyzing the Data with Hadoop, Scaling Out, Hadoop Streaming, **HDFS**: Design of HDFS, HDFS Concept, the command-Line Interface, Hadoop File-system, the Java interface, Data Flow, Data Integrity in Hadoop, Serialization, File based Data Structure. **Map Reduce**: Anatomy of a Map Reduce Job run, Failures, Job Scheduling, Shuffle and Sort, Task execution, Map Reduce Types and Formats, Map Reduce Features.	14
3	**Hadoop Operations** Setting up a Hadoop Cluster - Cluster Specification, Cluster Setup and Installation, Hadoop	06

	Configuration, Security in Hadoop.	
4	**Hadoop Eco System** **Pig:** Introduction to PIG, Execution Modes of Pig, Comparison of Pig with Databases, Grunt, Pig Latin, User Defined Functions, Data Processing operators. **Hive :** Hive Shell, Hive Services, Hive Metastore, Comparison with Traditional Databases, HiveQL, Tables, Querying Data and User Defined Functions	**06**
5	**NoSQL** Necessity of NoSQL, NoSQL business drivers; NoSQL case studies; NoSQL data architecture patterns: Key-value stores, Graph stores, Column family (Bigtable) stores, Document stores, Variations of NoSQL architectural patterns; Using NoSQL to manage big data, Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer-to-peer; Four ways that NoSQL systems handle big data problems	**07**
6	**Spark** Introduction to Data Analysis with Spark – Necessity for Spark, Unified Stack, Programming with RDD- Creating RDD, RDD Operations, Passing function to spark, working with key-value pair, Loading and saving your data, Spark Streaming.	**08**

Suggested Specification table with Marks (Theory): (For B. Tech only)

Distribution of Theory Marks					
R Level	**U Level**	**A Level**	**N Level**	**E Level**	**C Level**
10	20	20	10	0	0

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create and above Levels (Revised Bloom's Taxonomy)

Reference Books:

Sr No.	**Title of book /article**	**Author(s)**	**Publisher and details like ISBN**
1	Hadoop-The Definitive Guide	Tom White	O'Reilly
2	BIG Data and Analytics	Seema Acharya, Subhashini Chhellappan	Willey
3	Learning Spark	Jules S. Damji, Brooke Wenig, Tathagata Das & Denny Lee	O'Reilly
4	Understanding Big data	Chris Eaton, Dirk derooset	McGraw Hill
5	Learning Spark: Lightning-Fast Big Data Analysis	Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia	O'Reilly

Note: Students should refer to the latest editions of books

Course Outcomes:

Sr. No.	CO statements	Marks % weightage
CO-1	Understand the importance of Big Data, various sources of data and its Business Implications.	15%
CO-2	Manage and Analyze big data solutions using hadoop framework.	30%
CO-3	Develop Big Data Solutions using Hadoop EcoSystem.	30%
CO-4	Analyze data analysis with different techniques like Spark and its application.	25%

List of Open learning website:

- http://www.altova.com/xmlspy.html
- https://www.w3.org/RDF/

List of Experiments:

1. Prepare a document on big data containing below given topics:
    a. What is Big Data?
    b. Characteristics (Four V's) of Big Data
    c. Challenges of Big Data
    d. Applications of Big data
    e. What is Big Data Analytics?
    f. Types of Big Data Analytics
    g. How Big Data Analytics helps in development of smart city?
2. To install Hadoop framework, configure it and setup a single node cluster. Use web based tools to monitor your Hadoop setup.
3. With the help of Java shell commands perform the operation of how to include a file in HDFS.
    a. Show the working of different Java shell commands to work in HDFS.
4. Implement MongoDB CRUD operations.
    - Create a Student database and create a collection named Student within the Student database. Show the working of Create, Read, Update and Delete operations in the Student collection using

MongoDB.

5. Write a program to implement Flajolet Martin [FM] Algorithm to count distinct elements in the given data input stream.
6. To Install and Run Hive then use Hive to create, alter, and drop databases, tables. To create HDFS tables and load them in Hive and implement joining of tables in Hive.
7. To implement a word count application using the MapReduce API.