CS442: DATA SCIENCE AND ANALYTICS

Credits and Hours:

Teaching Scheme	Theory	Practical	Total	Credit
Hours/week	3	4	7	5
Marks	100	100	200	

A. **Pre-requisite courses:**

- Data Structures & Algorithm Design
- Database Management System
- Design & Analysis of Algorithms
- Computer Programming
- Engineering Mathematics

B. **Outline of the Course:**

Sr. No.	Title of the unit	Minimum number of hours
1.	INTRODUCTION TO DATA SCIENCE	04
2.	STATISTICAL INFERENCE	05
3.	DATA PRE-PROCESSING AND DATA VISUALIZATION	05
4.	INTRODUCTION TO MAP-REDUCE AND HADOOP ARCHITECTURE	05
5.	HDFS, HIVE AND HIVEQL, HBASE	10
6.	APACHE SPARK	06
7.	NoSQL	03
8.	DATA BASE FOR THE MODERN WEB	07

Total Hours (Theory): 45
Total Hours (Lab): 60
Total Hours: 105

C. **Detailed Syllabus:**

1. **INTRODUCTION TO DATA SCIENCE** **04 Hours 10%**

Introduction of data science and data analytics, Defining data science by its key components, Big Data and its importance, Four Vs, Drivers for Big data, Big data applications, Exploring Data Science in

Business, Applications in real-world

2. **STATISTICAL INFERENCE** **05 Hours** **08%**

Event Space, Random Variables and Probability Distributions

3. **DATA PRE-PROCESSING AND DATA VISUALIZATION** **05 Hours** **15%**

Dataset, Types of Dataset, Importance of Pre-processing the Data, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization and Concept Hierarchy Generation, Data visualization techniques

4. **INTRODUCTION TO MAP-REDUCE AND HADOOP ARCHITECTURE** **05 Hours** **12 %**

Big Data – Apache Hadoop & Hadoop EcoSystem, Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce, Data Serialization.

5. **HDFS, HIVE AND HIVEQL, HBASE** **10 Hours** **20 %**

HDFS-Overview, Installation and Shell, Java API; Hive Architecture and Installation, Comparison with Traditional Database, HiveQL Querying Data, Sorting And Aggregating, Map Reduce Scripts, Joins & Sub queries, HBase concepts, Advanced Usage, Schema Design, Advance Indexing, PIG, Zookeeper , how it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper.

6. **Apache SPARK** **06 Hours** **15%**

Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs, Machine Learning with MLlib.

7. **NoSQL** **03 Hours** **08 %**

What is it?, Where It is Used Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NewSQL

8. **Data Base for the Modern Web** **07 Hours** **12 %**

Introduction to MongoDB key features, Core Server tools, MongoDB through the JavaScript’s Shell, Creating and Querying through Indexes, Document-Oriented, principles of schema design, Constructing queries on Databases, collections and Documents,

MongoDB Query Language.

D. **Course Outcome (COs):**

After completion of the course, Students will be able to

CO1	Use an ethically responsible approach to evaluate and interpret data
CO2	Demonstrate expertise in statistical data processing
CO3	Use of various algorithms as well as mathematical and statistical models and optimization concepts to formulate and the use analyse data appropriately
CO4	Develop the ability to build and evaluate data-based models.
CO5	To learn difference between conventional SQL query language and NoSQL andMongoDBbasic concepts
CO6	Utilizing data science principles and approaches to solve real-life situational problems and effectively communicate them.

E. **Course Articulation Matrix:**

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CO1	-	1	-	-	-	2	2	3	2	-	3	2	3	1
CO2	3	1	2	-	2	-	-	-	1	-	1	1	2	2
CO3	3	2	3	2	3	-	3	-	1	2	3	3	3	2
CO4	1	-	3	1	1	-	2	-	2	-	1	1	3	3
CO5	-	3	1	1	3	-	-	-	-	-	1	-	2	-
CO6	1	2	3	3	1	3	3	2	3	3	3	2	2	3

F. **Recommended Study Material:**

❖ **Text book:**

1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, by EMC Education Services, Wiley, 2015.
2. Professional Hadoop Solutions By Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, Wiley, ISBN: 9788126551071, 2015.
3. Understanding Big data By Chris Eaton,Dirkderooset al. , McGraw Hill, 2012.
4. BIG Data and Analytics ,Sima Acharya, Subhashini Chhellappan, Willey
5. MongoDB in Action, Kyle Banker,PiterBakkum , Shaun Verch, Dream tech Press

6. HADOOP: The definitive Guide By Tom White, 4th Edition O Reilly 2012.
7. Big Data Analyticswith R and Haoop By VigneshPrajapati, Packet Publishing 2013.
8. Learning Spark: Lightning-Fast Big Data Analysis Paperback by Holden Karau, Apress

- **Reference book:**
  1. Big Data Analytics with Spark ByGuller, Mohammed,Apress
  2. Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications By Bart Baesens, Wiley Publication
  3. Hadoop in Practice by Alex Holmes, Manning Publication
- **Web material:**
  1. http://www.bigdatauniversity.com/
  2. https://sparkhub.databricks.com/resources/