

## Big Data

<b>Course Code</b>	19CS4701D	<b>Year</b>	IV	<b>Semester</b>	II
<b>Course Category</b>	<b>Program Elective-VI</b>	<b>Branch</b>	CSE	<b>Course Type</b>	Theory
<b>Credits</b>	3	<b>L-T-P</b>	3-0-0	<b>Prerequisites</b>	Data mining
<b>Continuous Internal Evaluation :</b>	30	<b>Semester End Evaluation:</b>	70	<b>Total Marks:</b>	100

Course Outcomes		
Upon successful completion of the course, the student will be able to		
<b>CO1</b>	Understand the basic concepts of big data	<b>L2</b>
<b>CO2</b>	Apply the concept of HDFS, Map reduce, Spark for storing and processing of Big data	<b>L3</b>
<b>CO3</b>	Apply Hive for working with Big data and formulate an efficient report	<b>L3</b>
<b>CO4</b>	Apply various analytics mechanisms to design a recommender system.	<b>L3</b>

Syllabus		
Unit No.	Contents	Mapped CO
<b>I</b>	<p><b>Meet Hadoop:</b> Data, Data Storage and Analysis, Querying All Your data, Beyond Batch, Comparison with Other Systems: Relational database Management Systems, Grid Computing, Volunteer Computing, A Brief History of Apache Hadoop.</p> <p><b>Map Reduce:</b> A Weather Dataset: Data Format, Analyzing the data with Unix Tools, Analyzing the Data with Hadoop: Map and Reduce, Java Map Reduce, Scaling Out: Data Flow, Combiner Functions, Running a Distributed Map Reduce Job.</p>	<b>CO1,CO2</b>
<b>II</b>	<p><b>The Hadoop Distributed Filesystem:</b> The Design of HDFS, HDFS Concepts, The Command Line Interface, Hadoop File systems, The Java</p>	<b>CO1,CO2</b>

	Interface, Data Flow, Parallel Copying with distcp.	
<b>III</b>	<b>Hive:</b> Hive Shell, An Example, Running Hive: Configuring Hive, Hive Services, The Metastore, Comparison with Traditional Databases: Schema on Read Versus Schema on write, Updates, transactions, and Indexes, SQL-on Hadoop Alternatives, <b>HiveQL:</b> Data Types, Operators and Functions and Tables: managed Tables and External Tables, Partitions and Buckets, Storage Formats, Importing Data, Altering Tables and Dropping Tables.	<b>CO1,CO3</b>
<b>IV</b>	<b>Spark:</b> Installing Spark, An Example: Spark Applications, Jobs, Stages, and Tasks, A Scala Standalone Application, A Java Example, A Python Example, Resilient Distributed Datasets: Creation, Transformations and Actions, Persistence, Serialization, Shared Variables: Broadcast Variables, Accumulators, Anatomy of a Spark Job: Run: Job Submission, DAG Construction, Task Scheduling, Task Execution.	<b>CO1,CO2</b>
<b>V</b>	<b>Use case Study:</b> Recommendation Systems: Introduction, A Model for Recommendation Systems, Collaborative Filtering System and Content Based Recommendations.	<b>CO1,CO4</b>

### Learning Resources

#### Text Book

1. Hadoop: The Definitive Guide, Tom White, Fourth Edition, 2015, O'Reilly.
2. Big Data Analytics, RadhaShankarmani, M Vijayalakshmi, Second Edition, 2017, Wiley

#### References

1. Hadoop Essentials: A Quantitative Approach, Henry H. Liu, First Edition, 2012, PerfMath Publishers
2. Big Data and Analytics, Seema Acharya, SubhashiniChellappan, First Edition, 2015, Wiley.
3. Big data analytics with R and Hadoop, VigneshPrajapati, First Edition, 2013, SPD.
4. Spark: The Definitive Guide :Big Data Processing Made simple, Bill Chambers and MateiZaharia, First Edition, 2018, O'Reilly

#### e-Resources and other Digital Material

1. <https://nptel.ac.in/courses/106/104/106104189/>
2. <https://www.coursera.org/specializations/big-data>
3. <https://www.edx.org/course/big-data-fundamentals>