20IT6404B BIGDATA

Course Category:	Program Elective-II	**Credits:**	3
Course Type:	Theory	**Lecture-Tutorial-Practice:**	3-0-0
Prerequisites:	20IT4304: Database Management Systems 20IT5404A: Data Mining	**Continuous Evaluation:**	30
		Semester end Evaluation:	70
		Total Marks:	100

Course Outcomes

Upon successful completion of the course, the student will be able to:

CO 1	Understand Big data characteristics, Hadoop, Hive, HDFS and Map Reduce architectures.
CO 2	Use Nosql Databases To Process Different Varieties of Data.
CO 3	Apply Pig Latin, Hive Scripts and Map Reduce programming on real time applications.
CO 4	Perform In Memory Data Analytics With Spark and Spark Streaming.

Contribution of Course Outcomes towards achievement of Program Outcomes (L-Low, M-Medium, H-High)

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2
CO 1	L												M	M
CO 2	M	M			H								H	M
CO 3	M		H		H								H	M
CO 4	L		M										M	M

Course Content

UNIT I

Introduction to Big Data:

Big Data-definition, Characteristics of Big Data (Volume, Variety, Velocity), Data in the Warehouse and Data in Hadoop, Importance of Big Data and Patterns for Big Data Development.

Introduction to Hadoop:

Data, Data Storage and Analysis, Comparison with Other Systems: RDBMS, A Brief History of Hadoop, Apache Hadoop and the Hadoop Ecosystem, Hadoop Releases.

UNIT II

NoSQL: Introduction to NOSQL, Types of NoSQL Databases, Advantages of NoSQL databases, SQL versus NoSql.

	NoSQL databases: Introduction to MongoDB, Data types in MongoDB, MongoDB query language. **Hadoop Distributed File System:** The Design of HDFS, HDFS Concepts, Blocks, Namenodes and Datanodes, Basic Filesystem Operations, Hadoop Filesystems, Interfaces, The Java Interface, Reading Data from a HadoopURL, Data Flow, Anatomy of a File Read and Anatomy of a File Write.
	UNIT III: **Map Reduce**–A Weather Dataset, Data Format, Analyzing the Data with Unix Tools, Analyzing the Data with Hadoop, Map and Reduce, Java Map Reduce, Scaling Out, Hadoop Streaming, Hadoop Pipes. **Pig**-Installation and Running of Pig, Execution Types, Running Pig Programs, Pig Latin Editors, Comparison with databases, Pig Latin, Functions, Data Processing Operators.
	UNIT IV: **Hive**-Installing Hive, An Example, Running Hive, Comparison with Traditional Databases, HiveQL, Tables, Querying Data. **Spark:** Introduction to data analytics with Spark, Spark Stack, Programming with RDDS, Working with key/value pairs and Spark SQL.
Text books and Reference books	**Text Book(s):** [1]. Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch, "Understanding Big Data Analytics for Enterprise Class Hadoop and Streaming Data" 1st Edition, TMH,2012. [2]. Tom White, Hadoop, "The Definitive Guide", 3rd Edition, O'Reilly Publications, 2012. [3]. Seema Acharya, Subhashini Chellappan, Big Data and Analytics, Wiley Publishers. **Reference Books:** [1]. Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, "Learning Spark: Lightning-Fast Big Data Analysis", O'Reilly Media, Inc.
E-resources and other digital material	[1]. Big Data Use cases for Beginners | Real Life Case Studies | Success Stories https://www.youtube.com/watch?v=HHR0-iJp2sM [2]. Alexey Grishchenko, Hadoop vs MPP, https://0x0fff.com/hadoop-vs-mpp/ [3]. Random notes on bigdata- SlideShare: Available www.slideshare.net/yiranpang/random-notes-on-big-data-26439474