Nirma University

Institute of Technology, School of Technology

MTech Computer Science and Engineering (Data Science)

Semester – II

L	T	P	C
3	0	2	4

Course Code	6CS271
Course Name	Big Data Systems

Course Learning Outcomes (CLOs):

At the end of the course, students will be able to

1. analyse the big data analytic techniques for business applications.
2. manage big data using different tools and frameworks.
3. design efficient algorithms for mining the data from large volumes.
4. implement the HADOOP and MapReduce technologies associated with big data analytics

Syllabus	**Teaching Hours**
Unit I	**5**
Introduction to Big Data: Introduction to Big Data Platform, Challenges of Conventional Systems, Intelligent Data Analysis, Nature of Data, Analytic Processes and Tools, Analysis vs Reporting, Modern Data Analytic Tools, Statistical Concepts: Sampling Distributions, Re-Sampling, Statistical Inference - Prediction Error	
Unit II	**10**
The Big data technology landscape : NoSQL, Types of No SQL databases, SQL Vs No SQL, why No SQL, Introduction to MongoDB, Data Types in MongoDB, CRUD, Practice examples, Apache Cassandra, Features of Cassandra, CRUD operations	
Unit III	**10**
Hadoop: History of Hadoop, The Hadoop Distributed File System, Components of Hadoop, Analysing the Data with Hadoop, Scaling Out, Hadoop Streaming, Design of HDFS, Java Interfaces to DFS Basics, Developing a Map Reduce Application, How Map Reduce Works, Anatomy of a Map Reduce Job Run, Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce Types and Formats, Map Reduce Features, Hadoop ecosystem.	

Unit IV **10**

Hadoop Environment: Setting up a Hadoop Cluster, Cluster Specification, Cluster Setup and Installation, Hadoop Configuration, Security in Hadoop, Administering Hadoop, HDFS, Monitoring, Maintenance, Hadoop benchmarks, Hadoop in the Cloud.

Unit IV **10**

Frameworks: Applications on Big Data Using Pig and Hive, Data Processing Operators in Pig, Hive Services, HiveQL, Querying Data in Hive, Fundamentals of HBase and ZooKeeper, IBM Info Sphere Big Insights and Streams, Visualizations, Visual Data Analysis Techniques, Interaction Techniques, Systems and Applications

Self-Study:

The self-study contents will be declared at the commencement of semester. Around 10% of the questions will be asked from self-study contents.

Laboratory Work:

Laboratory work will be based on above syllabus with minimum 6 experiments to be incorporated.

Suggested Readings^:

1. Michael Berthold, David J. Hand, Intelligent Data Analysis, Springer
2. Tom White, Hadoop: The Definitive Guide, Third Edition, O'reilly Media
3. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill Publishing
4. Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets, Cambridge University Press
5. Bill Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, John Wiley & sons
6. Glenn J. Myatt, Making Sense of Data, John Wiley & Sons
7. Pete Warden, Big Data Glossary, O'Reilly
8. Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, Second Edition, Elsevier
9. Da Ruan, Guoquing Chen, Etienne E. Kerre, Geert Wets, Intelligent Data Mining, Springer
10. Paul Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, James Giles, David Corrigan, Harness the Power of Big Data: The IBM Big Data Platform, Tata McGraw Hill Publications
11. Michael Minelli, Michele Chambers, Ambiga Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses, Wiley Publications
12. Zikopoulos, Paul, Chris Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Tata McGraw Hill Publications
13. Seema Acharya and Subhashini C, Big Data and Analytics, Wiley India

L=Lecture, T=Tutorial, P=Practical, C=Credit

^this is not an exhaustive list