

Program Code: ET203	
Subject Code: 1ET2030207	Subject Title: BIG DATA ANALYSIS (Elective – II)
Pre-requisite	-

Course Objective:

This course provides practical foundation level training that enables immediate and effective participation in big data projects. The course provides grounding in basic and advanced methods to big data technology and tools, including MapReduce and Hadoop and its ecosystem.

Teaching Scheme (Hours per week)				Evaluation Scheme (Marks)				
Lecture	Tutorial	Practical	Credit	Theory		Practical		Total
				University Assessment	Continuous Assessment	University Assessment	Continuous Assessment	
2	-	2	4	60	40	30	20	150

Subject Contents				
Sr. No	Topic	Total Hours	Weight (%)	
1	INTRODUCTION TO BIG DATA Introduction – distributed file system – Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications. Algorithms using map reduce, Matrix-Vector, Multiplication by Map Reduce.	4	16%	
2	INTRODUCTION HADOOP Big Data – Apache Hadoop & Hadoop Ecosystem – Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce - Data Serialization.	7	30%	
3	HADOOP ARCHITECTURE Hadoop Architecture, Hadoop Storage: HDFS, Common Hadoop Shell commands , Anatomy of File Write and Read., NameNode, Secondary NameNode, and DataNode, Hadoop MapReduce paradigm, Map and Reduce tasks, Job, Task trackers - Cluster Setup – SSH & Hadoop Configuration – HDFS Administering – Monitoring & Maintenance.	6	25%	
4	HADOOP ECOSYSTEM AND YARN Hadoop ecosystem components - Schedulers - Fair and Capacity, Hadoop 2.0 New Features NameNode High Availability, HDFS Federation, MRv2, YARN.	4	16%	
5	HIVE AND HIVEQL, HBASE Hive Architecture and Installation, Comparison with Traditional Database, HiveQL - Querying Data - Sorting And Aggregating.	3	13%	

Review Presentation: The student is expected to spend minimum 2 hours per week (as mentioned in the teaching and evaluation scheme) to refer at least two peer reviewed journal papers related to this domain/subject. The student is expected to identify issues/challenges and emerging trends in the domain/subject. Student is supposed to explore various video lectures (E.g. NPTEL) available in the domain/subject. Student is required to make a review-presentation on the work carried out for the same.

Recommended sites for journal papers are (1) dl.acm.org (2) springer.com (3) sciencedirect.com (4) elsevier.com (5) ieeexplore.ieee.org (6) scholar.google.co.in (7) scopus.com or others of similar repute.

Course Outcome:

Upon successful completion of this course the student would be able to:

- learn tips and tricks for Big Data use cases and solutions
- learn to build and maintain reliable, scalable, distributed systems with Apache Hadoop
- apply Hadoop ecosystem components

List of References:

1. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", Wiley, 2015.
2. Chris Eaton, Dirk deroos et al., "Understanding Big data", McGraw Hill, 2012.
3. Tom White, "HADOOP: The definitive Guide", O Reilly 2012

E-Resources:

1. <http://bigdatauniversity.com/courses/?cat=124>
2. <http://www.coreservlets.com/hadoop-tutorial/>

List of Experiments:

1. Master the concepts of HDFS and MapReduce framework.
2. Understand Hadoop 2.x Architecture.
3. Setup Hadoop Cluster and write Complex MapReduce programs.
4. Learn data loading techniques using Sqoop and Flume.
5. Perform data analytics using Pig, Hive and YARN.
6. Implement HBase and MapReduce integration.
7. Implement Advanced Usage and Indexing.
8. Schedule jobs using Oozie.
9. Implement best practices for Hadoop development.