Undergraduate Course Description Document

1.	Department offering the course	Computer Science and Engineering
2.	Course Code	CSF443
3.	Course Title	Big Data Analytics
4.	Credits (L:T:P:C)	2:0:1:3
5.	Contact Hours (L:T:P)	2:0:2
6.	Prerequisites (if any)	
7.	Course Basket	

Course Summary

To learn the need for Big Data Analytics, and to acquire modern tools to implement in real life applications.

Course Objectives

Understanding the fundamentals of various big data analysis techniques, Hadoop structure, environment and framework**.**

Course Outcomes

- Understand the need and process of data analysis.
- Learn the different component of Hadoop Ecosystem.
- Understand the Map Reduce and the use of Apriori and Fp-Growth.
- Learn to analyse the data using R.
- Analyse different software for processing Big Data.

Curriculum Content

UNIT 1: INTRODUCTION TO BIG DATA AND HADOOP [5]

Types of Digital Data, Introduction to Big Data, Big Data Analytics, Analytic Processes and Tools, Analysis vs Reporting, Statistical Concepts: Sampling Distributions, Re-Sampling, Statistical Inference, Prediction Error, Modern Data Analytic Tools - History of Hadoop, Apache Hadoop, Analysing Data with Unix tools, Analysing Data with Hadoop, Hadoop Streaming, Hadoop Echo System, IBM Big Data Strategy.

UNIT 2: HADOOP DISTRIBUTED FILE SYSTEM (HDFS) [5]

The Design of HDFS, HDFS Concepts, Command Line Interface, Hadoop file system interfaces, Data flow, Data Ingest with Flume and Scoop and Hadoop archives, Hadoop I/O: Compression, Serialization, Avro and File-Based Data structures.

UNIT 3: MAP REDUCE [5]

Anatomy of a Map Reduce Job Run, Failures, Job Scheduling, Shuffle and Sort, Task Execution, Map Reduce Types and Formats, Map Reduce Features. Mining Frequent Item sets :- Market Based Model, Apriori Algorithm, FP-Growth.

UNIT 4: HADOOP ECO SYSTEM [5]

Pig: Introduction to PIG, Execution Modes of Pig, Comparison of Pig with Databases, Grunt, Pig Latin, User Defined Functions, Data Processing operators. Hive: Hive Shell, Hive Services, Hive Metastore, Comparison with Traditional Databases, HiveQL, Tables, Querying Data and User Defined Functions. Hbase: HBasics, Concepts, Clients, Example, Hbase Versus RDBMS. Big SQL: Introduction.

UNIT- 5: DATA ANALYTICS WITH R [6]

Overview of R programming language, Regression Modelling, Multivariate Analysis. Machine Learning: Introduction, Supervised Learning, Unsupervised Learning, Collaborative Filtering. Big Data Analytics with BigR. Machine learning tools: Spark & SparkML, H2O, Azure ML

Textbook(s)

1. **Intelligent Data Analysis**, Michael Berthold, David J. Hand,2/e, Springer, 2015.
2. **Mining of Massive Datasets**, Anand Raja Raman and Jeffrey David Ullman,2/e, Cambridge University Press, 2012.
3. **Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics,** Bill Franks, 1/e, John Wiley & sons, 2012.
4. **Hadoop: The Definitive Guide**, Tom White Third Edition, O’reillyMedia,2012.

Reference Books

1. **Making Sense of Data,** I, Glenn J. Myatt, 2/e, John Wiley & Sons, 2014
2. **Big Data Glossary**, Pete Warden,1/e, O’Reilly, 2011.
3. **Data Mining Concepts and Techniques**, Jiawei Han, Micheline kamber, 2/e, Elsevier, Reprinted 2015.

List of Experiments:

s. No.	Title of experiment
1	Installation of Hadoop.
2	Directory Management Tasks in Hadoop a. Create a directory in HDFS b. List the Contents of directory c. Remove a directory in HDFS
3	File Management Tasks in Hadoop a. Upload and download a file in HDFS b. See Contents of a File. c. Remove a file in HDFS.
4	File Transfer in Hadoop a. Copy a file from Source to destination. b. Move file from Source to Destination.
5	Word Count Map Reduce program to understand MAP Reduce Paradigm.
6	Weather Report POC-Map Reduce Program to analyse time-temperature statistics and generate report with max/min temperature.
7	Implementing Matrix Multiplication with Hadoop Map Reduce.
8	Pig, Latin Scripts to sort, Group, Join Project and Filter the data.

9	Introduction to Weka tool to process data.
10	Use R to process data and visualize it using ggplot2

Tools/Software for experiments: Hadoop

Teaching and Learning Strategy

All materials (ppts, assignments, labs, etc.) will be uploaded in Moodle. Teaching of students will be conducted through power point lectures, tutorials, short classroom exercises.

Benchmarking:

1. **Columbia University, New York** (https://www.ee.columbia.edu/~cylin/course/bigdata/)
2. **The Graduate Institute Geneva** **(https://www.karstendonnay.net/download/spring2018/Syllabus_MINT-078.pdf)**
3. **NSUT Delhi** **(http://www.nsit.ac.in/static/documents/IS.pdf)**
4. **IIIT Delhi** **(https://www.iiitd.ac.in/academics/courses/institute#CSE510A)**