Syllabus for B.Tech (CSE) III YEAR II SEM
Big Data Analytics
(Professional Elective – III)

L	T	P/D	C
3	0	0	3

Code: 9FC15
Prerequisites: Database Management Systems

Course Objective:
Explore the fundamental techniques and principles in achieving big data analytics with stream processing.

Course Outcomes:
After completion of this course student will be able to:

1. Comprehend the fundamentals of big data analytics using Hadoop to solve the real life problems. [L2]-U1
2. Outline the concepts of map reduce, Hive in big data environment and differentiate NoSQL and SQL databases. [L2, L4]-U2,U3
3. Develop the algorithms to process big data using Apache Spark Low Level API. [L3]-U4,U5
4. Make use of Stream Processing techniques to develop social media applications. [L3]-U6

UNIT– I:
Introduction to Big Data: Big Data Analytics, Characteristics of Big Data – The Four Vs, importance of Big Data, Different Use cases, Data-Structured, Semi-Structured, Un-Structured

Introduction to Hadoop and its use in solving big data problems. Comparison Hadoop with RDBMS, Brief history of Hadoop, Apache Hadoop EcoSystem, Components of Hadoop, The Hadoop Distributed File System (HDFS):, Architecture and design of HDFS in detail, Working with HDFS (Commands)

UNIT-II
Anatomy of Hadoop map-reduce (Input Splits, map phase, shuffle, sort, combiner, reduce phase) (theory)

Hive: Introduction to Hive, data types and file formats, HiveQL data definition(Creating Databases and Tables), HiveQL for Data loading, HiveQL data manipulation, Logical joins, Window functions, Optimization, Table partitioning, Bucketing, Indexing, Join Strategies.

UNIT-III
SQOOP: Introduction to SQOOP, SQOOP imports: From Database to HDFS/Hive, SQOOP exports: From HDFS/Hive to Database, Incremental imports

NoSQL & HBase: Overview, HBase architecture, CRUD operations

UNIT-IV
SPARK Basics: History of Spark, Spark Architecture, Spark Shell, working with RDDs in Spark: RDD Basics, Creating RDDs in Spark. RDD Operations. Passing Functions to Spark, Transformations and Actions in Spark, Spark RDD Persistence

Working with Key/Value Pairs: Pair RDDs, Transformations on Pair RDDs, Actions

Available on Pair RDDs

UNIT-V

Structured API : DataFrames, SQL : Overview of Structured Spark Types, Schemas, Columns and Expressions, DataFrame Transformations, Working with different types of data, Aggregations- Aggregation Functions, Grouping, User-Defined Aggregation Functions, ,Joins-Inner Joins, Outer Joins, Processing CSV Files, JSON Files, Text Files and Parquet Files, Spark SQL

UNIT-VI

Spark streaming: Stream Processing Fundamentals, Structured Streaming Basics - Core Concepts, Structured Streaming in Action, Transformations on Streams, Input and Output (Kafka)
Case study: Twitter Stream processing application

TEXT BOOKS:

1. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012
2. SPARK: The Definitive Guide, Bill Chambers &MateiZaharia, O'Reilley, 2018 Edition

REFERENCES:

1. "Hadoop Operations", O'Reilley, Eric Sammer,2012
2. "ProgrammingHive",O'Reilley,E.Capriolo,D.Wampler,andJ.Rutherglen,2012
3. "HBase: The Definitive Guide", O'Reilley, Lars George,2011
4. Big Data, Big Analytics: Emerging, Michael Minnelli, Michelle Chambers, and Ambiga Dhiraj