4/4 B.Tech. FIRST SEMESTER

CS7T3 **BIG DATA CONCEPTS** **Credits: 4**

(Analytics)

Required

Lecture: 4 periods/week **Internal assessment: 30 marks**

Tutorial: 1 period /week **Semester end examination: 70 marks**

Course Context and Overview: This course explores a range of the most relevant topics that pertain to contemporary analysis practices, technologies and tools for Big Data environments.

Prerequisite: Data Mining, Distributed Systems

Objectives:

After successful completion of this course, Student will be able to

1. Understand the history of Hadoop and the associated computing techniques.
2. Analyze the Weather Dataset with Unix Tools and Hadoop Tools.
3. Analyze the Hadoop Distributed File system.
4. Analyze the Avro Data Serialization System.
5. Evaluate Map Reduce Application development and working process.
6. Analyze the types and formats of Map Reduce.
7. Analyze the Features of Map Reduce.

Learning Outcomes:

After successful completion of this course, Student would

1. Analyze the data with Hadoop framework
2. Explain HDFS concepts, interfaces, and basic file system operations
3. Understand the fundamentals of i/o in hadoop
4. Devlop and impliment Map reduce applications on hadoop
5. Explore Map reduce types and input formats and output formats

UNIT I

Introduction to Hadoop: Data, Data types, Storage and Analysis, Relational Database Management System, Grid Computing, Volunteer Computing, A Brief History of Hadoop, Apache Hadoop and the Hadoop Ecosystems.

UNIT II

Map Reduce: A Weather Dataset: Data Format, Analyzing the data with Unix Tools, Analysing the Data with Hadoop: MapReduce, Java MapReduce, Scaling Out: Data Flow, Combiner Function,s Running a Distributed Map Reduce Job, Hadoop Streaming: Ruby, Python, Hadoop Pipes, Compiling and Running.

UNIT III

The Hadoop Distributed Filesystem: The Design of HDFS, HDFS Concepts, The Command_Line Interface, Hadoop Filesystems, The Java Interface, Data Flow, Data Ingest with Flume and Sqoop, Parallel Copying with distcp and Hadoop Archieves.

UNIT IV

Hadoop I/o: Data Integrity, Compression, Serialization, Avro(Data Serialization System): Avro Data Types and Schemes, In-Memory Serialization and Deserialization, Avro Datafiles, Interoperatbility, Schema Resolution, Sort order, Avro Map Reduce, Sorting using Avro Map Reduce, Avro Map Reduce in other Languages, File-Based Data Structures.

UNIT V

Developing a Map Reduce Application: The Configuration API: Setting up the Development Environment, Writing a Unit Test with MRUnit, Running Locally on Test Data, Running on a cluster, Tuning a Job, Map Reduce Workflows.

UNIT VI

How Map Reduce Works: Anatomy of a Map Reduce Job Run, Failures, Job Scheduling, Shuffle and Sort, Task Execution.

UNIT VII

Map Reduce Types and Formats: Map Reduce Types, Input Format: Input Splits and Records, Text Input, Binary Input, Multiple Inputs, Database Input and Output, Output Formats: Text Output, Binary Output, Multiple Outputs, Lazy Output, Database Output.

UNIT VIII

Map Reduce Features: Counters: Built-i9n Counters, User-Defined Java Counters, User-defined Streaming Counters, Sorting: Preparation, Partial Sort, Total Sort, Secondary Sort, Joins: Map-Side Joins, Reduce-Side Joins, Side Data Distribution: Using the Job Configuration, Distributed Cache, Map Reduce Library Classes.

Learning Resources

Textbook:

Hadoop: The Definitive Guide, Tom White, 3rd Edition (2012), O'Reilly(SPD).

Reference:

Hadoop Essentials: A Quantitative Approach, Henry H. Liu, 1st Edition (2012), PerfMath Publishers.