**MDS571: BIG DATA ANALYTICS**

**Total Teaching Hours for Trimester: 75**

**No of hours per week: 7(4+3)**

**Max Marks: 100 Credits: 4**

**Course Type: Major**

**Course Description**

The subject is intended to give the knowledge of Big Data evolving in every real-time

application and how they are manipulated using the emerging technologies. This course

breaks down the walls of complexity in processing Big Data by providing a practical

approach to developing Java applications on top of the Hadoop platform. It describes the

Hadoop architecture and how to work with the Hadoop Distributed File System (HDFS).

**Course Outcomes:** Upon completion of the course students will be able to

|  |  |
| --- | --- |
| **No.** | **Course Outcomes** |
| **CO1** | Understand the Big Data concepts in real time scenario |
| **CO2** | Identify different types of Hadoop architecture |
| **CO3** | Demonstrate an ability to use Hadoop framework for processing Big Data for Analytics |
| **CO4** | Analyze the Big data under Spark architecture |
| **CO5** | Demonstrate the programming of Big data using Hive and Pig environments |

**Unit 1 Introduction**

Concepts of Data Analytics: Descriptive, Diagnostic, Predictive, Prescriptive analytics - Big

Data characteristics: Volume, Velocity, Variety, Veracity of data - Types of data: Structured,

Unstructured, Semi-Structured, Metadata - Introduction to Hadoop Scaling - Distributed

Framework -Hadoop v/s RDBMS-Brief history of Hadoop.

**Teaching Hours: 15**

**Unit 2 Big Data Architecture**

Standard Big data architecture - Big data application - Hadoop framework - HDFS Design

goal - Master Slave architecture - Block System - Read-write Process for data - Installing

HDFS - Executing in HDFS: Reading and writing Local files and Data streams into HDFS

- Types of files in HDFS - Strengths and alternatives of HDFS - Concept of YARN.

Apache Hadoop Moving Data in and out of Hadoop Understanding inputs and outputs of

MapReduce - Problems with traditional large-scale systems-Requirements for a new

approach.

**Teaching Hours: 15**

**Unit 3 Parallel Processing with MapReduce**

Introduction to MapReduce - Sample MapReduce application: Wordcount - MapReduce Data

types and Formats - Writing MapReduce Programming - Testing MapReduce Programs -

MapReduce Job Execution - Shuffle and Sort - Managing Failures - Progress and Status

Updates. MapReduce Programs: Using languages other than Java with Hadoop, Analyzing a

large dataset.

**Teaching Hours: 15**

**Unit 4 Hive and Pig**

Hive Architecture - Components - Data Definition - Partitioning - Data Manipulation -

Joins, Views and Indexes - Hive Execution - Pig Architecture - Pig Latin Data Model -

Latin Operators - Loading Data - Diagnostic Operators - Group Operators - Pig Joins - Row

Level Operators - Pig Built-in function - User defined functions - Pig Scripts.

**Teaching Hours: 15**

**Unit 5 Stream Processing with Spark**

Stream processing Models and Tools - Apache Spark - Spark Architecture: Resilient

Distributed Datasets, Directed Acyclic Graph - Spark Ecosystem - Spark for Big Data

Processing: MLlib, Spark GraphX, SparkR, SparkSQL, Spark Streaming - Spark versus

Hadoop . PySpark + NumPy + SciPy, Code Optimization.

**Teaching Hours: 15**

**Essential Reading**

1. Anil Maheshwari (2020). *Big Data*. 2nd Edition. McGraw Hill Education Pvt Ltd.

2. S Chandramouli, Asha A George, C R Rene Robin,,D Doreen H Miriam,J JasmineC M,

*Big Data Analytics*,University Press India Ltd., 2024

**Recommended Reading**

1. Thomas Erl, Wajid Khattak and Paul Buhler (2016). *Big Data Fundamentals: Concepts,*

*Drivers and Techniques*. Service Tech Press.

2. Julián Luengo, Diego García-Gil, Sergio Ramírez-Gallego, Salvador García, Francisco

Herrera (2020). *Big Data Preprocessing: Enabling Smart Data*. Springer Nature

Publishing.

3. Seema Acharya, Subhasini Chellappan (2019), *Big Data and Analytics*. 2nd Edition,

Wiley India Pvt Ltd