**CA20319A** **(3L + 1 T hrs./week)**

## BIG DATA AND ITS APPLICATIONS IN CLOUD

No. of questions to be set: Total 5 questions will be given from Unit I & Unit II.
No. of questions to be answered: All questions to be answered.

**Objectives:** This course provides an overview of the concept of Big Data, Machine Learning and Cloud Computing. It provides different aspect of knowledge discovery for Big data and various data mining process for climate data and multilevel text mining. It also present the Hadoop framework for distributed computing and MapReduce for parallel processing and stream processing with Spark for data management.

**Pre-requisites:** Fundamentals of data mining concepts.

**Course Outcomes (CO):**

| | |
|---|---|
| CO1 | Students will able to understand the concepts knowledge discovery from big data. |
| CO2 | Students will be able to understand the concept of machine learning. |
| CO3 | Students will be able handle the programming language for distributed computing. |

### UNIT-I

**Introduction to Big Data [6 hrs]**
Understanding Big Data, capturing Big Data: Volume of Data, Velocity of Data, variety of Data, Veracity of Data. Benefitting from Big Data, Management of Big Data, Organising and Analysing Big Data.

**Machine Learning and Incremental Learning with Big data [4 hrs]**
Machine learning concepts, Big data and machine learning, Incremental learning, Incremental learning for knowledge building, Incremental technique to handle big data, Applications

**Knowledge discovery for big data [4 Hrs]**
Opinion Mining: Aspect and entity extraction, Data Mining for climate data , Multilevel Text Mining.

**Introduction to cloud computing [6 hrs]**
Need for cloud computing, Business and IT perspectives, Benefits and challenges of cloud computing, cloud module: Public, private, Hybrid, community cloud. Cloud application architecture: Grid computing, transaction computing. Cloud computing architecture, cloud services: Database as a service, infrastructure as a service, platform as a service.

### UNIT-II

**Distributed computing using Hadoop [6 hrs]**
Hadoop Framework, HDFS Design goals, Master-slave architecture, Block system, sequence files, YARN

**Parallel processing with Map Reduce [7 hrs]**
MapReduce overview, Sample Map Reduce Application: WordCount. MapReduce Programming: Data Types and Format, Writing MapReduce programming, Testing MapReduce programs. MapReduce Jobs Execution: Managing Failures, process & status updates. Hive Language and Pig language.

**Stream processing with Spark [7 hrs]**
Spark Architecture, Resilient Distributed Datasetrs (RDDs), Directed Acylic GraPH(dag), Spark Ecosystem. Spark for Big Data processing: MLlib, Spark GraphX, Spark R, Spark SQL, Spark Streaming. Spark vs Hadoop.

**Text Books:**

1. Anil Maheshwari, Big Data, McGraw-Hill Education
2. M.N Rao, Cloud Computing, PHI learning Pvt. Ltd.


**Reference Books:**

1. Wesley W. Chu Data Mining & Knowledge Discovery for Big Data: Methodologies, challenges and opportunities, Springer Science and Business Media.
2. VenkatAnkam, Big data Analytics, Packt Publishing Ltd.
3. Kim H. Pries, Robert Dunnigan, Big Data Analytics: A Practical Guide for Managers, CRC Press.