18MC301BIG DATA ANALYTICS

Course Description and Objectives:

This course gives an overview of Big Data, i.e. storage, retrieval and processing of big data. In addition, it also focuses on the "technologies", i.e., the tools/algorithms that are available for storage, processing of Big Data. It also helpsa student to perform a variety of "analytics" on different data sets and to arrive at positive conclusions.

Course outcomes:

The student will be able to:

- Analyze the theoretical issues involved in Big Data system design such as curse of dimensionality, Data classification and predictive analytics etc...
- Get familiarization with major approaches in Big Data Analytics.
- Hands-on Big Data tools like Hadoop, Pig & Hive.
- Apply machine learning techniques on given data sets.

Skills:

- Design classification models for various standard datasets and user datasets.
- Develop clustering techniques and association rules for large standard datasets and user datasets.
- Analyse large scale data using MAPREDUCE programming which includes JAVA and HADOOP frameworks.
- Analyse large scale data using PIG and HIVE.

Activities:

- Logistic Regression using MENARCHE dataset.
- Logistic Linear Regression using TITANIC dataset.
- Decision Tree & Cross Validation using IRIS dataset.
- Random Forest & Cross Validation using IRIS dataset.

Syllabus

UNIT – 1 **14 Hours**

ESSENTIALS OF BIG DATA AND ANALYTICS: Data, Characteristics of data and Types of digital data, Sources of data, Working with unstructured data, Evolution and Definition of big data, Characteristics and Need of big data, Challenges of big data; Overview of business intelligence, Data science and Analytics, Meaning and Characteristics of big data analytics, Need of big data analytics, Classification of analytics, Challenges to big data analytics, Importance of big data analytics, Basic terminologies in big data environment.

UNIT – 2 **12 Hours**

HADOOP :Introducing Hadoop, Need of Hadoop, limitations of RDBMS, RDBMS versus Hadoop, Distributed computing challenges, History of Hadoop , Hadoop overview, Use case of Hadoop, Hadoop distributors, HDFS (Hadoop Distributed File System) , Processing data with Hadoop, Managing resources and applications with Hadoop YARN (Yet another Resource Negotiator), Interacting with Hadoop Ecosystem.

UNIT – 3 **12 Hours**

MAPREDUCE PROGRAMMING:Introduction , Mapper, Reducer, Combiner, Partitioner, Searching, Sorting, Compression, Real time applications using MapReduce, Data serialization and Working with common serialization formats, Big data serialization formats.

UNIT – 4 **12 Hours**

HIVE**:** Introduction to Hive, Hive architecture, Hive data types,Hive file format, Hive Query Language (HQL), User-Defined Function (UDF) in Hive;

UNIT – 5 **10 Hours**

PIG: The anatomy of Pig , Pig on Hadoop, Pig Philosophy, Use case for Pig; ETL Processing , Pig Latin overview , Data types in Pig , Running Pig , Execution modes of Pig, HDFS commands, Relational operators, Piggy Bank , Word count example using Pig.

Text Book:

SeemaAcharya, SubhashiniChellappan, "Big Data Analytics", 1st Edition, Wiley, 2015.

Reference Books:

1. Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, "Professional Hadoop Solutions", 1st Edition, Wrox, 2013.
2. Chris Eaton,DirkDerooset. al., "Understanding Big data", Indian Edition, McGraw Hill, 2015.
3. Tom White, "HADOOP: The definitive Guide", 3rd Edition, O Reilly, 2012.
4. VigneshPrajapati, "Big Data Analytics with R and Hadoop", 1st Edition, Packet Publishing Limited, 2013.