III Year II Semester	L	T	P	C
Code: 20DS6660	4	0	0	4

LARGE SCALE DATA PROCESSING (Honors)

Course Objectives:

The course should enable the students to:

1. Understand different techniques of systems modeling, clustering, virtualization.
2. Be familiar with concepts of cloud platform architecture.
3. Be exposed to data analytics big data principles and map reduce.
4. Implementation of Beginning Apache Pig Big Data Processing.
5. Learning Spark Lightning-Fast Data Analytics

Course Outcomes:

After completion of this course, students able to

1. Summarize the techniques of systems modeling, clustering, virtualization
2. Understand the cloud platform architecture.
3. Distinguish big data principles.
4. Apply Apache Pig Big Data Processing.
5. Understand the Spark Lightning

UNIT –I

Systems modeling, Clustering and virtualization: Scalable Computing over the Internet, Technologies for Network based systems, System models for Distributed and Cloud Computing, Software environments for distributed systems and clouds, Performance, Security And Energy Efficiency

UNIT-II

Cloud Platform Architecture: Cloud Computing and service Models, Architectural Design of Compute and Storage Clouds, Public Cloud Platforms, Inter Cloud Resource Management, Cloud Security and Trust Management. Service Oriented Architecture, Message Oriented Middleware.

UNIT-III

Big Data Principles and Paradigms: Real-Time Analytics, Big Data Analytics for Social Media, Deep Learning and Its Parallelization, Characterization and Traversal of Large Real-World Networks, Database Techniques for Big Data**,** Map Reduce- Map Reduce, Hadoop, Map Reduce Algorithms- Hadoop usage patterns, Map Reduce Examples, Map Reduce Jobs.

UNIT-IV

Beginning Apache Pig Big Data Processing: Pig- Apache pig, running pig, Pig Latin, Pig Latin Data Structures, Pig Example, Word count in pig, pig workflow, Advantages and disadvantages of Pig.

UNIT-V

Learning Spark Lightning-Fast Data Analytics: Spark- Introduction to Apache Spark: A Unified Analytics Engine, Downloading Apache Spark and Getting Started, Apache Spark's Structured APIs, Spark SQL and Data Frames: Introduction to Built-in Data Sources, Spark SQL and Data Frames: Interacting with External Data Sources, Spark SQL and Datasets.

Text Books:

1. Distributed and Cloud Computing, Kai Hwang, Geoffry C. Fox, Jack J. Dongarra MK Elsevier.
2. Cloud Computing, Theory and Practice, Dan C Marinescu, MK Elsevier.
3. Big Data Principles and Paradigms Edited by RajkumarBuyya The University of Melbourne and Manjra soft Pty Ltd, Australia Rodrigo N. Calheiros The University of Melbourne, Australia Amir VahidDastjerdi The University of Melbourne, Australia

Reference Books:

1. Beginning Apache Pig Big Data Processing Made Easy BalaswamyVaddeman
2. Hadoop: The Definitive Guide Tom White

Reference Links:

1. https://youtu.be/r5k-_RLIpuA.
2. https://youtu.be/NzZXz3fJf6o