SYLLABUS

Course Code	Course/Subject Name	Credits
CSDLO7032	Big Data Analytics	4

Course Objectives (CO):

1. To provide an overview of an exciting growing field of big data analytics.
2. To introduce programming skills to build simple solutions using big data technologies such as MapReduce and scripting for NoSQL, and the ability to write parallel algorithms for multiprocessor execution.
3. To teach the fundamental techniques and principles in achieving big data analytics with scalability and streaming capability.
4. To enable students to have skills that will help them to solve complex real-world problems in for decision support.
5. To provide an indication of the current research approaches that is likely to provide a basis for tomorrow's solutions.

Course Outcomes : Students should be able to -

1. Understand the key issues in big data management and its associated applications for business decisions and strategy.
1. Develop problem solving and critical thinking skills in fundamental enabling techniques like Hadoop, Mapreduce and NoSQL in big data analytics.
2. Collect, manage, store, query and analyze various forms of Big Data.
3. Interpret business models and scientific computing paradigms, and apply software tools for big data analytics.
4. Adapt adequate perspectives of big data analytics in various applications like recommender systems, social media applications etc.
5. Solve Complex real world problems in various applications like recommender systems, social media applications, health and medical systems, etc.

Pre-requisites : Some prior knowledge about Java programming, Basics of SQL, Data mining and machine learning methods would be beneficial.

Module	Detailed Contents	Hrs.
01	**Introduction to Big Data and Hadoop** 1.1 Introduction to Big Data, 1.2 Big Data characteristics, Types of Big Data, 1.3 Traditional vs. Big Data business approach, 1.4 Case Study of Big Data Solutions. 1.5 Concept of Hadoop 1.6 Core Hadoop Components; Hadoop Ecosystem **(Refer Chapters 1 and 2)**	06

02	**Hadoop HDFS and MapReduce** 2.1 Distributed File Systems : Physical Organization of Compute Nodes, Large-Scale File-System Organization. 2.2 MapReduce : The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures. 2.3 Algorithms Using MapReduce : Matrix-Vector Multiplication by MapReduce, Relational-Algebra Operations, Computing Selections by MapReduce, Computing Projections by MapReduce, Union, Intersection, and Difference by MapReduce 2.4 Hadoop Limitations **(Refer Chapter 3)**	10
03	**NoSQL** 3.1 Introduction to NoSQL, NoSQL Business Drivers, 3.2 NoSQL Data Architecture Patterns: Key-value stores, Graph stores, Column family (Bigtable)stores, Document stores, Variations of NoSQL architectural patterns, NoSQL Case Study 3.3 NoSQL solution for big data, Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models: master-slave versus peer-to-peer; NoSQL systems to handle big data problems. **(Refer Chapter 4)**	06
04	**Mining Data Streams** 4.1 The Stream Data Model : A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing. 4.2 Sampling Data techniques in a Stream 4.3 Filtering Streams : Bloom Filter with Analysis. 4.4 Counting Distinct Elements in a Stream, Count-Distinct Problem, Flajolet-Martin Algorithm, Combining Estimates, Space Requirements 4.5 Counting Frequent Items in a Stream, Sampling Methods for Streams, Frequent Itemsets in Decaying Windows. 4.6 Counting Ones in a Window: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows. **(Refer Chapter 5)**	12
05	**Finding Similar Items and Clustering** 5.1 Distance Measures : Definition of a Distance Measure, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance. 5.2 CURE Algorithm, Stream-Computing , A Stream-Clustering Algorithm, Initializing & Merging Buckets, Answering Queries **(Refer Chapters 6 and 7)**	08
06	**Real-Time Big Data Models** 6.1 PageRank Overview, Efficient computation of PageRank : PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector. 6.2 A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering. 6.3 Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities in a social graph. **(Refer Chapters 8, 9 and 10)**	10

Module - 1

Syllabus :

Introduction to Big Data, Big Data characteristics, Types of Big Data, Traditional vs. Big Data business approach, Case Study of Big Data Solutions, Concept of Hadoop, Core Hadoop Components, Hadoop Ecosystem

Module - 2

Syllabus :

Distributed File Systems : Physical Organization of Compute Nodes, Large-Scale File-System Organization, MapReduce : The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping With Node Failures, Algorithms Using MapReduce : Matrix-Vector Multiplication by MapReduce, Relational-Algebra Operations, Computing Selections by MapReduce, Computing Projections by MapReduce, Union, Intersection, and Difference by MapReduce, Hadoop Limitations

Module - 3

Syllabus :

Introduction to NoSQL, NoSQL Business Drivers, NoSQL Data Architecture Patterns: Key-value stores, Graph stores, Column family (Bigtable)stores, Document stores, Variations of NoSQL architectural patterns; NoSQL Case Study, NoSQL solution for big data, Understanding the types of big data problems; Analyzing big data with a shared-nothing architecture; Choosing distribution models : master-slave versus peer-to-peer; NoSQL systems to handle big data problems.

Chapter 4 : NoSQL 4-1 to 4-27

Module - 4

Syllabus :

The Stream Data Model: A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing, Sampling Data techniques in a Stream, Filtering Streams: Bloom Filter with Analysis, Counting Distinct Elements in a Stream, Count-Distinct Problem, Flajolet-Martin Algorithm, Combining Estimates, Space Requirements, Counting Frequent Items in a Stream, Sampling Methods for Streams, Frequent Itemsets in Decaying Windows, Counting Ones in a Window: The Cost of Exact Counts, The Datar-Gionis-Indyk-Motwani Algorithm, Query Answering in the DGIM Algorithm, Decaying Windows.

Chapter 5 : Mining Data Streams 5-1 to 5-17

Module - 5

Syllabus :

Distance Measures : Definition of a Distance Measure, Euclidean Distances, Jaccard Distance, Cosine Distance, Edit Distance, Hamming Distance., CURE Algorithm, Stream-Computing, A Stream-Clustering, Algorithm, Initializing & Merging Buckets, Answering Queries

Module - 6

Syllabus :

PageRank Overview, Efficient computation of PageRank : PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector, A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering, Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities in a social graph

Chapter 10 : Mining Social Network Graph 10-1 to 10-14

❑❑❑