15B1WCI833: Big Data Analytics

Course Credit: 3
Semester: VIII

Introduction

This course introduces basic technology (algorithms, architectures, systems) and advanced research topics in connection with large-scale data management and information extraction techniques for big data. The course will start by introducing the fundamentals of Big data and cover modern distributed database systems and algorithms and Big data systems adopted in industry and science applications. Distributed storage and parallel processing and architectures that support data analytics will be examined, and students will learn how to implement a distributed data processing system. The course will also cover critical topics in mining and knowledge discovery of big data, with applications in social analytics, cyber security, and information networks, among others that are already in public eye.

Course Objectives (Post-conditions)

Knowledge objectives:

1. Describe how to represent data and information for processing.
2. Evaluate different methodologies for effective application of data mining.
3. Define “speed-up” and explain the notion of an algorithm’s scalability.
4. Explain basic statistical concepts and their areas of application.
5. Parallelize an algorithm by applying data-parallel decomposition.
6. Decompose a problem using map and reduce operations.
7. Discuss the importance of elasticity and resource management in cloud computing.

Application objectives:

1. Understand and apply the Big Data Flow to actual projects.

2. Being able to describe and apply the Data Analytics lifecycle to Big Data projects and lead other team members in the process.

Expected Student Background (Preconditions)

1. Good knowledge of Statistics
2. Working knowledge of databases

Good knowledge of data structures and algorithms

Topics Outline:

Serial Number	Topics	Hours
1	Introduction to Big Data: Big data time line, Why this topic is relevant now? Is big data fad? Where using big data makes a difference? Introduction to statistical modeling and machine learning, Ordinary data processing versus big data processing: Challenges and opportunities	3
2	Map Reduce and the New Software Stack: Distributed File Systems, Map Reduce, Algorithms Using Map Reduce, Complexity Theory for Map Reduce	3
3	Mining Data Streams: The Stream Data Model, Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments and Windowing, Decaying Windows	5
4	Link Analysis: Page Rank and Efficient Computation of Page Rank, Topic-Sensitive Page Rank, Link Spam, Hubs and Authorities	5
5	Frequent Item sets from Big	7

	Data: The Market-Basket Model, Market Baskets and the A-Priori Algorithm, Handling Larger Datasets in Main Memory, Limited-Pass Algorithms, Counting Frequent Items in a Stream	
6	Clustering for Big Data: Introduction to Clustering Techniques, Hierarchical Clustering, Clustering in Non-Euclidean Spaces, Clustering for Streams and Parallelism	8
7	Mining Social Network Graphs: Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities, Partitioning of Graphs, Finding Overlapping Communities, Neighborhood Properties of Graphs	6
8	Recommendation Systems: A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering and Dimensionality Reduction	5
	Total Lectures	42

References

1. Anand Rajaraman and Jeffery David Ullman, Mining of Massive Datasets, Cambridge University Press, 2012
2. Jared Dean, Big Data, Data Mining and Machine Learning, Wiley Big data Series, 2014
3. Judith Hurwitz, Alan Nugent, Fern Halper and Marica Kaufman, Big Data for Dummies, Wiley Press, 2013

Evaluation Scheme:

S. No.	Examination	Marks
1	T1	15
2	T2	25
3	T3	35
4	* Internal Marks	25

*Internal Marks Breakdown:

Assignments 9 marks (3x3)

Quizzes 12 marks (3x4)

Regularity 4 Marks