Course Name:	Map Reduce and Big Data	**Approval: 9th Senate Meeting**
Course Number:	CS 561	
Credits:	3-0-0-3	
Prerequisites:		
Intended for:	UG/PG	
Distribution:	Elective	
Semester:		

Preamble:

In the current computing world, big data problems play a big role and every CS graduate must be aware of these problems and how best to solve them. This course is designed to meet that need and prepare the student for real-world computing. This course is also suitable for preparing research scholars intending to work in the area of big data, information retrieval and machine learning.

Course Outline:

This course aims to provide the students with a thorough understanding of the MapReduce paradigm and its application to Big Data problems. The course will provide a hands-on programming experience and teach the student to think in terms of Map and Reduce to solve any kind of large scale problem that deals with huge data and needs distributed fault-tolerant computation.

Unit 1 and 2 will be delivered through lectures while Unit 3 will be delivered by a combination of lectures, student presentations and programming projects/demonstrations.

Pre-requisites: Programming in Java and/or Python and/or Scala.

Modules:

- **Unit 1 (3 hrs):** Introduction to Big Data, the MapReduce paradigm and programming model. The MapReduce framework and its benefits. Cloud computing and MapReduce. Open source MapReduce frameworks – Hadoop, Shark, Mrjob etc. - comparison and benefits of each such framework.
- **Unit 2 (3 hrs):** Thinking in the MapReduce way via simple problems that serve as a building block for larger problems – Matrix-Vector multiplication, Matrix multiplication, Relational algebra – selections, projections, union, intersection, difference, natural join, grouping and aggregation. Complexity analysis of MapReduce algorithms.
- **Unit 3 (36 hrs): Applying MapReduce to different Big Data contemporary problem areas.**
    - Similarity computations
    - Clustering algorithms – K-means, CURE, ...
    - Web crawling and indexing

- Web-scale graph algorithms – PageRank, HITS, ...
- Recommendation Systems – content-based filtering, collaborative filtering, dimensionality reduction.
- Text advertising on the web – AdWords.
- Social networks and their analysis.
- Large-scale Machine Learning – Perceptrons and Support Vector machines.
- Statistical Machine Translation.
- Market Basket analysis.
- Mining Data Streams.

Textbooks:

1. Jimmy Lin and Chris Dyer - *Data-Intensive Text Processing with MapReduce* – Morgan and Claypool.
2. Rajaraman and Ullman - *Mining of massive Datasets* – Cambridge Univ. Press