جـــامـــعـــة مـحـمـد بـــن زايـــد
للـــذكـــاء الاصـــطـــنـــاعـــي
MOHAMED BIN ZAYED UNIVERSITY
OF ARTIFICIAL INTELLIGENCE

Core Courses Syllabi

DS702 - Big Data Processing

Title	Big Data Processing
Code	DS702
Loading	4 Credit-hours
Prerequisites	• Databases • Proficiency in Java or Python
Catalog Description	This course is an introductory course on big data processing, which is the process of analyzing and utilizing big data. The course involves methods at the intersection of parallel computing, machine learning, statistics, database systems, etc.
Goal	The aim of this course is to provide students with the comprehensive understanding of the academic and industrial development of big data processing foundations and techniques. Students will understand the basic concepts of parallel computing, big data, MapReduce, Hadoop, etc. and will be able to develop advanced skills to solve practical big data processing problems.
Contents	This course introduces the basic concepts, principles, methods, implementation techniques, and applications of MapReduce, with a focus on **(I)** the MapReduce paradigm and **(II)** the Hadoop framework.
Recommended Textbooks	1. S. K. Prasad, A. Gupta, A. Rosenberg, A. Sussman, and C. Weems, Topics in Parallel and Distributed Computing, Springer International Publishing, 2018. 2. J. Lin and C. Dyer, Data-intensive Text Processing with MapReduce, Morgan & Claypool Publishers, 2010. 3. T. White, Hadoop: The Definitive Guide, O'Reilly Media, 2009.
Recommended References & Supplemental Material	Relevant research papers, technical reports, and surveys for each topic, where needed, are identified in the teaching plan ahead.

Teaching Week	Topics
1	**Introduction to Big Data and Parallel Computing** **Lecture** • Course orientation • Course overview • Basic concepts of big data • Basic concepts and taxonomy of parallel computing **Lab** • Instructor-led discussion on the topics taught in the week
2	**Introduction to MapReduce** **Lecture** • History of MapReduce • Basic idea of MapReduce • Programming model and design methodology of MapReduce • Basic applications of MapReduce **Lab** • Instructor-led discussion on the topics taught in the week
3	**Architecture of Google MapReduce** **Lecture** • Basic architecture and working principle of Google MapReduce • Basic architecture and working principle of GFS • Distributed, structured Bigtable **Lab** • Instructor-led discussion on the topics taught in the week
4	**Architecture of Hadoop MapReduce** **Lecture** • Basic architecture and working principle of Hadoop • Basic architecture and working principle of HDFS • Basic Hadoop programming **Lab** • Instructor-led discussion on the topics taught in the week

Teaching Week	Topics
5	**Hadoop Installation and Hadoop programming** **Lecture** • Single node Hadoop • Multi-node Hadoop • Remote job submission and execution • Hadoop programming **Lab** • Hadoop installation • Implement the WordCount algorithm (Assignment 1)
6	**Algorithm Design using Hadoop MapReduce (1)** **Lecture** • Computational problems that can be handled by Hadoop MapReduce • Sorting using Hadoop MapReduce **Lab** • Sort large volume of data using Hadoop MapReduce (Assignment 2)
7	**Algorithm Design using Hadoop MapReduce (2)** **Lecture** • Create an inverted index for a large corpus of documents • The word co-occurrence algorithm • Application of the algorithms in analyzing patent documents **Lab** • Create an inverted index to support effective and efficient searching of documents (Assignment 3) • Midterm Exam preparation
8	**Introduction to Hadoop HBase** **Lecture** • The working principle of Hadoop HBase • HBase programming **Lab** • HBase programming (Assignment 4)

Teaching Week	Topics
9	**Introduce to Hadoop Hive** **Lecture** • Basic structure of Hadoop Hive • The working principle of Hadoop Hive • Hive programming **Lab** • Hive programming (Assignment 5)
10	**Advanced MapReduce Programming (1)** **Lecture** • User-defined functions • Hadoop I/O • Composite key-value pairs and their usage • Partitioner and combiner **Lab** • Instructor-led discussion on the topics taught in the week
11	**Advanced MapReduce Programming (2)** **Lecture** • Iterative MapReduce algorithms • Chaining MapReduce jobs in Hadoop • Linking multiple data sources in Hadoop • Accessing relational databases **Lab** • Instructor-led discussion on the topics taught in the week
12	**Search Engine Algorithms based on MapReduce** **Lecture** • Introduction to PageRank • MapReduce PageRank **Lab** • Implement the MapReduce PageRank algorithm and calculate the PageRank values of Wikipedia webpages (Assignment 6)

Teaching Week	Topics
13	**MapReduce-based Data Mining (1)** **Lecture** • Introduction to clustering • Clustering using MapReduce **Lab** • Implement the MapReduce k-means algorithm (Assignment 7)
14	**MapReduce-based Data Mining (2)** **Lecture** • Introduction to classification • Classification using MapReduce **Lab** • Implement the MapReduce kNN algorithm (Assignment 8)
15	**MapReduce-based Data Mining (3)** **Lecture** • Introduction to frequent itemset mining • The PSON algorithm **Lab** • Review and final exam preparation