ACM Data Science Task Force Course Example

The Basis of Big Data Computation
Harbin Institute of Technology, Harbin
Hongzhi Wang, Xianmin Liu

Knowledge Areas that contain competencies (knowledge, skills, and dispositions) covered in the course

Knowledge Area	Total Number of Contact Hours
Continuing professional development	2
Algorithms	12
Data structures	8
Big data computing architectures Parallel computing frameworks Distributed data storage Parallel programming	22
Techniques for Big Data applications	4
Continuing professional development	2

Where does the course fit in your undergraduate Data Science curriculum?

Big Data Computing Foundation is a course orientation course in computer science/software engineering, and it is a limited course for computer majors (including computer science and technology, Internet of Things engineering, bio-information science, information security), and software engineering. The first elements are data structure and algorithm, computer system, computer network, database system, software engineering, program language design. The aim is to enable students to understand the basic knowledge of big data direction, to lay the thinking foundation for students to study big data direction to solve large-scale data problems in the future, to understand the current research methods and technology application system of big data science, to understand the problem solving ideas and ability training ideas of big data disciplines, and then to understand the professional requirements and competency requirements in the future work of big data disciplines, and to improve the core competitiveness of students seeking jobs in big data disciplines.

What is covered in the course?

Big Data Computing Foundation is a course orientation course in computer science/software engineering, and it is a limited course for computer majors (including computer science and technology, Internet of Things engineering, bio-information science, information security), and software engineering. The first courses are data structure and algorithm,

computer system, computer network, database system, software engineering, program language design. The aim is to enable students to understand and understand the basic knowledge of big data direction, to lay the thinking foundation for students to study big data direction to solve large-scale data problems in the future, to understand and understand the current research methods and technology application system of big data science, to understand the problem solving ideas and ability training ideas of big data disciplines, and then to understand the professional requirements and competency requirements in the future work of big data disciplines, and to improve the core competitiveness of students seeking jobs in big data disciplines.

What is the format of the course?

The total number of hours in this course is 72 hours, of which 48 hours are taught in the large class and 24 hours are in the experiment. This is a course that combines theory and application. Classwork provides much of the content and expectations of the course. Curriculum related experiments are designed to allow students to better experience the practical application of the theories they have learned.

How are students assessed?

This course has a score of 100. It consists of:
Open big homework (30%): Using the submission principle, the teacher evaluates the excellent big homework will be displayed in the class. Big Job Topics: Design an application that applies big data technology based on current big data applications, such as industrial, medical, and financial applications, requiring coverage from requirements analysis to system design, and writing course reports.
Interactive experiments (20%): Students evaluate each other on the completion of their experiments. Based on whether the experiment meets the requirements and whether the experimental quality meets the expected level, the experimental topic is the construction and application of Hadoop/Spark system.
Final exams (50%): Covers all teaching content, with big data algorithms (30%), big data structure (15%), big data computing systems (30%), big data management systems (15%), and big data applications (10%).

Course tools and materials

This course does not require any materials, the teacher will send the materials to everyone when teaching. In addition, the following two textbooks can be used as teaching reference for students to supplement reading:
1. Wang Hongzhi. Big Data Algorithms . China Machine Press, 2015.
2. Lu Jiaheng. Hadoop's actual combat. China Machine Press, 2011.

Why do you teach the course this way?

This course is mainly in the form of classroom face-to-face, interspersed with small class discussions, flip classes, and experiments. Classroom face-to-face teaching is mainly taught by teachers, as the most common form of teaching, can be more comprehensive and systematic transfer of the main knowledge points to everyone. Small class discussion can promote communication and exchange between students, the classroom face-to-face process of problems to focus on solving, timely answer questions and puzzles, so that students have a clearer understanding of knowledge, while in the process of discussion can deepen their understanding of the corresponding knowledge points, in order to achieve the new effect of warm knowledge. Flip the classroom by the teacher before the class to produce a teaching

video about each knowledge point, students in the extracurricular to complete the task of watching the video, according to the teacher's explanation of knowledge for independent learning, in the classroom teachers and students, students and students to communicate with each other, solve problems. Students change from passive learning to active knowledge-seeking. Stimulate students' motivation and initiative in learning. Exercise students' ability to learn independently and think independently. Doing experiments can enhance the practical ability to operate, put the knowledge learned into use, not only on paper, general talk, the understanding of knowledge points more in place, more thorough.

Body of Knowledge coverage

KA	Sub-domain	Competencies Covered	Hours
PR	Continuing professional development	1. Understand the differences and differences between big data computing and ordinary computing. 2. Understand big data research ideas. 3. Understand big data computing thinking	2
PDA	Algorithms	Understand the basic concepts and parallel computing models of parallel algorithms, and evaluate parallel algorithms. Grasp the design ideas of Map-Reduce algorithm and understand examples of Map-Reduce algorithm. Understand the external memory algorithm design model and master the external memory sequencing design ideas Understand I/O search structure and external memory graph processing algorithm. Grasp the basic data model of data stream, understand sampling algorithm, data stream counting problem, online aggregation problem. Understand the basic concepts of crowdsourcing and the design ideas of platforms and crowdsourcing algorithms, and understand the big data processing	12

		algorithms based on crowdsourcing, including crowdsourcing connection and entity recognition algorithms.	
PDA	Data structures	1. Understand the peculiarities and principles of big data data structure. 2. Master the principles of Bloom filter and minhash. 3. Can apply big data data structure to solve practical problems.	8
BDS	Big data computing architectures Parallel computing frameworks Distributed data storage Parallel programming	1. Understand the concept of data-intensive distributed computing systems. 2. Have a deep understanding of distributed file system and distributed computing theory. 3. Master the principles of distributed systems, and have a deep understanding and knowledge of storage, queues, computing, and cluster management. 4. Have application development experience based on Hadoop and Spark, and be familiar with its architecture and operating principles. 5. Familiar with the operation of big data clusters, and have the operation and maintenance capabilities of big data clusters. 6. Have in-depth experience and understanding of Map-Reduce principles. 7. Understand the principles and application scenarios of parallel database systems and distributed database systems.	22

		8. Master the concepts and design ideas of NoSQL and NewSQL. 9. Will use relevant knowledge to analyze and solve practical problems.	
BDS	Techniques for Big Data applications	1. Understand data-driven business concepts and have strong logical analysis capabilities. 2. Have a deep understanding of the future big data application field, and will carry out related design and application. 3. Cultivate basic big data awareness and broaden the application knowledge in the field of big data.	4