**交叉信息研究院研究生新课程申请表**

**IIIS New Graduate Course Application Form**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 中文课名  Course Name in Chinese | | | | 大数据系统基础(A) | | | | |
| 英文课名  Course Name in English | | | | Big Data Systems (A) | | | | |
| 授课语言  Language | 中文 | | 开课学期  Semester | | 秋Fall  春Spring | | 学分  Credit | 3 |
| 讲课教师  Lecturer | 徐葳、黄隆波 | | 首次开课学年  First Course Given in （year） | | | 2014 | 考核方式  Assessment Method | 1.考试Exam |
| 开课方式  Course Format | | 1. Lecture讲授为主 | | | | | | |
| 申请理由与预期达到的目标（本课程对培养研究生有何重要作用，通过本课程的学习能使学生在哪些知识和能力方面得以提高，你拟采取何种办法达到此目标）：Application Reason & Expected Goals （course’s significance to graduates cultivation; abilities achieved after taking this course for the students; how would you achieve this goal）  大数据是一门交叉学科。本门课程作为大数据方向硕士两门信息类公共基础课之一，主要希望通过对一系列与大数据相关的计算机系统知识的学习和动手实践，让学生理解大数据分析系统的软硬件架构，目前可用的工具、技术上存在的挑战。  考虑到学生群体中有相当多的经管、社科及公共管理专业学生，我们注重讲授大数据系统工具的概念、应用场景以及商业价值。培养学生在工作中应用与选择适合大数据工具的能力，同时也为有志于继续深入学习大数据专业课程的学生创造基础。 | | | | | | | | |
| 与同类课程比较（与校内及世界一流大学已开设的同类课程相比，本课程在教学内容、教学方式等方面有哪些特色或特点）Comparison to other similar courses (Compared to other similar courses in Tsinghua or world top-level universities, the distinct features of this course in teaching contents, teaching methods, etc.)  同期开设的大数据系统基础B是面向信息类学生的公共课。更注重讲授系统工具的原理与实现，培养学生设计和实现新的系统工具的能力。要求学生具备计算机编程、数据库以及操作系统等本科计算机专业知识。此门课程不需要。因此受众更广。信息类学生如果不熟悉系统方向的基础知识，也可以两门课同步选修。  现有课程《大数据平台系统》是与工业界（目前是微软研究院）合办的讲座课程，面向信息类研究型硕士和博士，重点讲授大数据的前沿研究内容。本门课程在知识结构上更成系统，而对先修知识要求更低，对动手实验能力要求更多，因此更适于做必修课。  即将开设的《大数据系统科研专题？？》是更深入的大数据科研内容的介绍，重点介绍高级的系统设计方法和应用，面向信息类博士生。是建立在大数据系统基础II基础上的提高课。  大数据系统是大数据技术的重要组成部分，国际各学校的数据科学相关专业，以及计算机科学的相关专业均有开设，面向的对象包括计算机、经管等各学科，侧重点均有不同。我们充分考虑了我们自身学生的水平，以及与《大数据分析学》课程讲授的统计与算法的衔接，设计了此门课程的教学计划。 | | | | | | | | |
| 预备知识 Preliminary Knowledge Required  具备计算机基础知识，最好熟悉一门程序设计语言，学过或同时选修《大数据分析学》课程中讲授的概率统计相关内容。 | | | | | | | | |
| 适用学科、专业及选课限制Applicable Subjects or Majors  全校任选课。大数据方向硕士公共基础课（必修） | | | | | | | | |
| 主要教材（原则上不超过3本）Teaching Materials (No more than 3) | | | | | | | | |
| 主要参考书Reference Books | | | | | | | | |
| 课程内容简介Course Contents  本门课程通过对一系列与大数据相关的计算机系统知识的学习和动手实践，让学生理解大数据分析系统的软硬件架构，目前可用的工具、技术上存在的挑战。本课程是全校公共课，我们注重讲授大数据系统工具的概念、应用场景以及商业价值。培养学生在工作中应用与选择适合大数据工具的能力，同时也为有志于继续深入学习大数据专业课程的学生创造基础。  本课程的主要内容包括：  云计算、大数据文件系统、主要大数据计算框架、关系和非关系数据库、文本检索、大规模机器学习框架、流数据处理、深度学习系统、推荐算法、数据集的清洗和装载、大数据中的安全与隐私保护、以及典型的大数据应用。 | | | | | | | | |
| 教学大纲Teaching Syllabus   |  |  | | --- | --- | | 1  绪论 | 大数据系统行业概览及课程概述  大数据的价值链及代表公司或技术  大数据的成功商业案例分析  计算机组成原理基础，处理器、内存、I/O的概念  网络和网络协议栈 | | 2  云计算 | 云计算  认识超大规模数据中心的主要优势  理解云计算的模式和主要技术、对大数据的影响  案例分析：亚马逊的EC2平台  案例分析：百度的大数据系统架构和软件定义一切的理念  案例分析：OpenStack与开放源代码对大数据行业的影响 | | 3  文件存储 | 一致性理论，CAP定理  存储，文件系统  I/O与存储的主要思想，磁盘，SSD等的特性  ZooKeeper, Chubby 一致性  案例分析：GFS | | 4  计算框架 | 分布式系统的主要理念  Hadoop  Spark  GraphLab | | 5  关系数据库 | 数据库基本概念  关系数据库的模型，关系代数  关系数据库的设计，ER模型  SQL  事务处理的概念 | | 6  非关系数据库 | 按列存储的概念  NoSQL数据库，MongoDB的技术和典型应用  图数据库 GraphDB，Neo4j 的典型应用  Key-Value 存储和典型系统，Cassandra  数据库形式的优劣比较 | | 7  文本检索 | 倒排表和文本索引  TF/IDF  分布式索引与搜索引擎 | | 8  分析库 | MLlib  Apache Mahout  推荐系统  PageRank算法 | | 9  流数据处理和其他处理框架 | Storm  Spark Streaming  Hive  Dremel  Shark  资源统一管理和调度 Mesos | | 10  深度学习 | Deep Learning 的基本思想  Deep Learning的GPU实现  Deep Learning 的分布式实现 | | 11  数据集的清洗和装载 | 了解ETL  Flume，  结构化数据装载Sqoop | | 12  安全和隐私保护 | 大数据中的安全和隐私问题  大数据隐私保护的主要思路和技术途径  保护隐私的数据查询 | | 13  可视化 | 大数据可视化的方法  案例分析：Tableau | | 14  典型应用 | 物联网大数据应用  智慧城市机器应用 | | | | | | | | | |
| 院长审批 Endorsement by IIIS Dean    ( ) 同意Approve ( ) 不同意Not Approve  签字Signature:  日期Date: | | | | | | | | |
| 学位委员会审批意见Endorsement by the Academic Committee  ( ) 同意Approve ( ) 不同意Not Approve  签字Signature:  日期Date: | | | | | | | | |