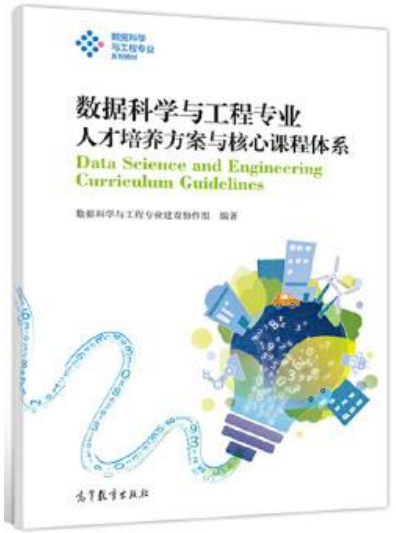
分布式计算系统

徐 辰
cxu@dase.ecnu.edu.cn

華東師範大學

课程名称

- 研究生
  - 2018、2019年：大数据处理系统
  - 2020年：大规模数据处理系统
  - 2021年：分布式计算系统



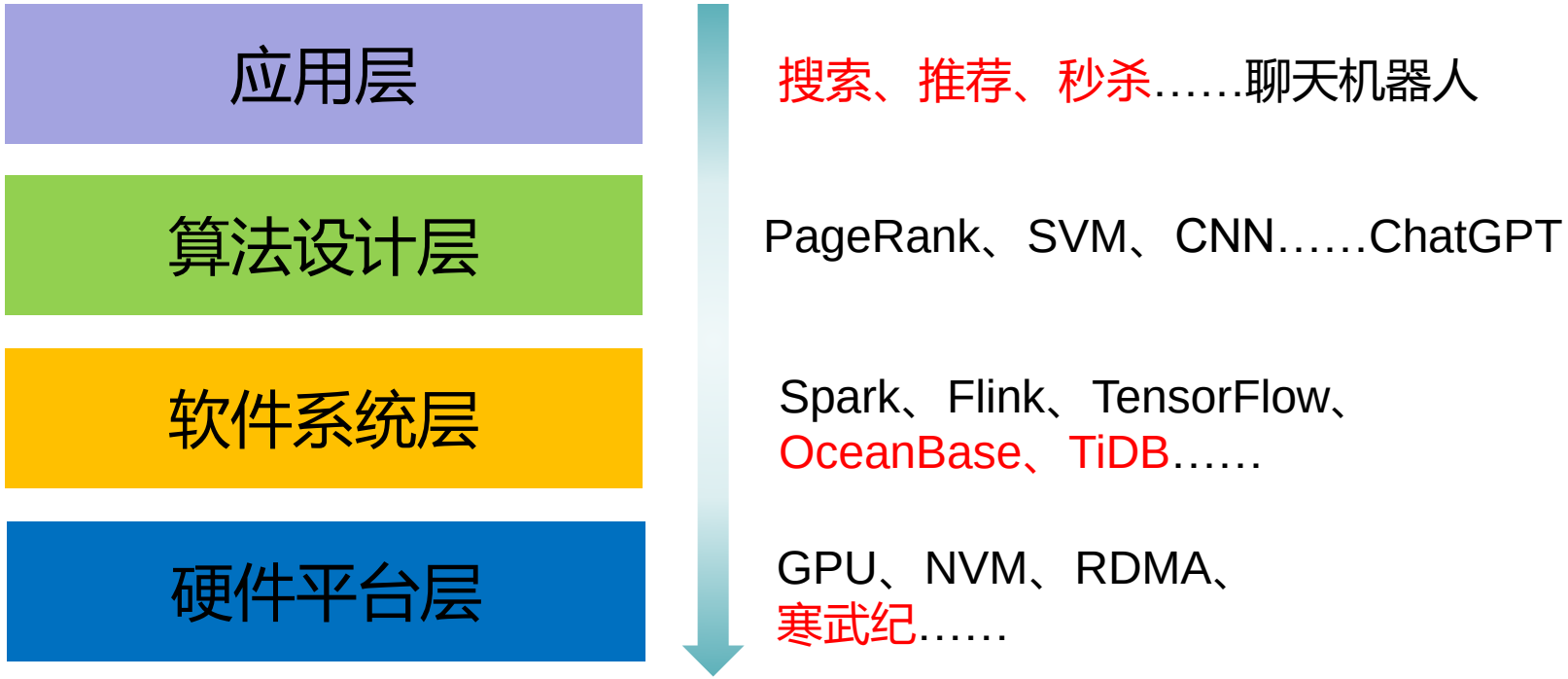
- 本科生
  - 2018、2019年：分布式模型与编程
  - 2021、2022年：分布式编程模型与系统
  - 2023年：分布式计算系统

课程背景

- 大数据处理系统 → 分布式计算系统
  - Hadoop、Spark、Flink等
  - “大数据”的涵义过于宽泛
- 其它类似课程/教材
  - 英文论文的翻译：不同论文的体系可能不一致
  - 针对某一系统的工具手册：时效性
- 本课程/教材
  - 强调系统设计、原理、编程的结合

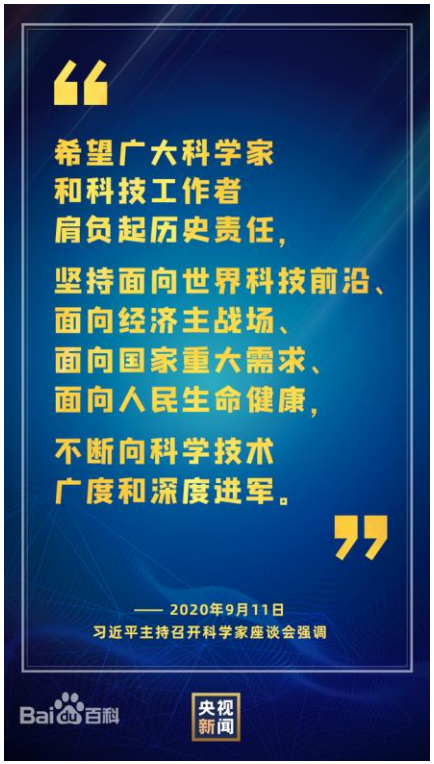
课程目的

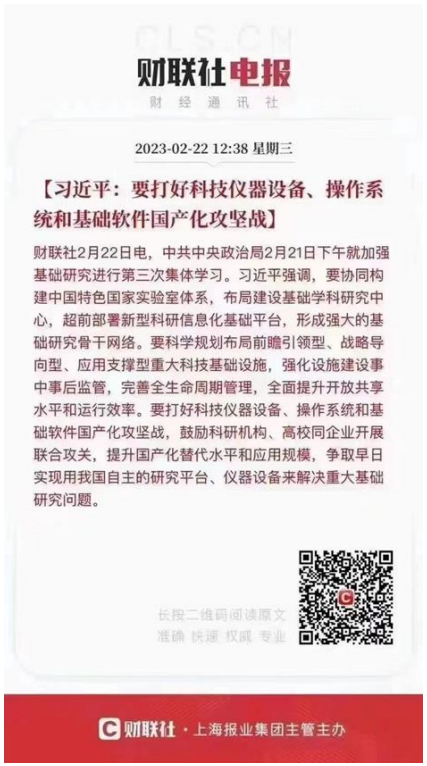
- 培养“系统思维”


应用层
算法设计层
软件系统层
硬件平台层
搜索、推荐、秒杀……聊天机器人
PageRank、SVM、CNN……ChatGPT
Spark、Flink、TensorFlow、
OceanBase、TiDB……
GPU、NVM、RDMA、
寒武纪……

课程目的

- 基础软件系统是国家重大战略需求，支撑国民经济和社会发展





课程内容

- When：背景
- Why：设计
- What：架构
- How：
  - 原理：系统层面
  - 编程：用户层面

课程荣誉

- 2023年上海高校市级重点课程

- 2023年上海高校市级一流本科课程

序号	开课单位	课程名称	课程负责人
1	计算机科学与技术学院	离散数学	卢兴见
2	政治与国际关系学院	国际政治概论	叶淑兰
3	教育学部	创新创业与教育实践变革	董辉
4	物理与电子科学学院	大学物理实验	尹亚玲
5	哲学系	中国近代哲学史	刘梁剑
6	音乐学院	和声与多声音乐	郑艳
7	软件工程学院	数据结构与算法	王丽萍
8	马克思主义学院	毛泽东思想和中国特色社会主义理论体系概论	陈红娟
9	传播学院	论辩与说服	甘莅豪
10	生态与环境科学学院	生态学研究方法与实验设计	邓泓
11	哲学系	科学史与科学方法	朱晶
12	教育学部	学前儿童数学教育	黄瑾
13	生态与环境科学学院	种群生态学	李德志
14	法学院	环境法	王欢欢
15	设计学院	插画设计：手绘	陈澜
16	通信与电子工程学院	低维传感器件与表征测试	吴幸
17	美术学院	漆艺	马俊营
18	生命科学学院	植物学	田怀珍
19	社会发展学院	社会学原著选读	刘拥华
20	体育与健康学院	体育概论	李琳
21	中国语言文学系	中国现代文学史	凤媛
22	数学科学学院	大学数学	贾挚
23	数据科学与工程学院	分布式计算系统	徐辰

课程安排(tentative)

周	周二	周四
1	绪论	实验一：准备工作
2	HDFS	实验二：Hadoop 1部署
3	HDFS	实验三：Hadoop 2部署
4	MapReduce设计思想与架构	实验三：Hadoop 2部署
5	MapReduce工作原理	实验四：MapReduce 2编程
6	MapReduce编程	清明节
7	MapReduce编程	实验四：MapReduce 2编程
8	Spark设计思想与架构	实验五：Spark部署
9	Spark工作原理	实验五：Spark部署

课程安排(tentative)

周	周一	周四
10	Spark编程	实验六：Spark编程
11	Spark编程、习题课	实验六：Spark编程
12	Yarn	实验七：Spark+Yarn
13	Flink设计思想与架构	实验八：Flink部署
14	Flink工作原理	实验八：Flink部署
15	Flink编程	实验九：Flink编程
16	习题课	实验九：Flink编程
17	总结、答疑	实验十：Flink+Yarn
18	考试	

课程安排

- 理论课程：每周2学时，紧跟节奏思考
  - 设计思想：为什么？
  - 系统架构：是什么？不同系统的联系与区别
  - 编程思路：不是教API
- 线上线下结合：
  - 在线慕课：课前预习，课后复习（是什么）
  - 课堂授课：增加问题讨论（为什么）

课程安排

- 实践课程：每周2学时
  - 开源系统部署：保持耐心、坑很多，不要奢望照着实验说明就能一步到位
  - 基本编程开发、代码调试：
    - 使用Java开发，动手能力强的自学Scala
    - 熟练使用IntelliJ IDE、maven
    - 遇到错误，设置断点调试，先动脑，再用搜索工具

课后训练

- 理论复习
  - 多思考，不要死记硬背
- 动手编程是最关键的
  - 上机作业：在线提交
  - 增强调试代码的能力

课程成绩评定

- 平时成绩：50%
  - 考勤：10%
  - 课堂讨论：15%
  - 编程作业（2+1次）：30%
  - 上机实验（10个实验共3组）：45%
- 期末成绩：50%
  - 期末考试：100%

课程要求

- 考勤：无故缺勤扣10分，扣完为止
- 课堂讨论：请积极参与课堂讨论
- 编程作业：3-4周时间内完成
- 上机实验：（本机+云环境）
  - 单机实验(90%)：独立完成，助教检查
  - 分布式实验（bonus，10%）：非必须内容，1-4人合作完成，完成后请主动找助教登记
  - 三组实验报告：实验2-4、实验5-7、实验8-9

实验报告

- 报告内容：需要记下来以后参考的内容
  - 不是抄实验说明，仅需记录你踩过什么坑，即：遇到了什么问题？是如何解决的？
  - 每个实验的报告不少于1页但不超过2页，每组实验报告的篇幅控制在6-8页
- 命名格式：学号-姓名-第X组实验-版本Y.pdf
  - X取值范围[1, +∞)
  - 例如：10061-张三-第1组实验-版本1.pdf
- 注意：网上提交，规定的deadline一般是实验课结束一周左右，不交0分，迟交扣分

实验验收及报告提交总体安排

周	实验内容		验收截止时间	实验报告提交截止时间
1		实验一：准备工作	无需验收	
2	第1组实验	实验二：Hadoop 1部署	3.13	4.24
3		实验三：Hadoop 2部署	3.27	
4				
5		实验四：MapReduce 2编程	4.17	
6				
7				
8	第2组实验	实验五：Spark部署	5.1（根据放假情况调整）	5.29
9				
10		实验六：Spark编程	5.15	
11				
12		实验七：Spark+Yarn	5.22	
13	第3组实验	实验八：Flink部署	6.5	6.23
14				
15		实验九：Flink编程	6.19	
16				
17		实验十：Flink+Yarn	无需验收	

编程作业(tentative)

- 每次5道题目，每题10分

难度系数	权重	备注
无	10%	赠送
易	20%	
中	30%	
中	30%	
难	10%	可选

提高作业

- 一个简易MapReduce系统的实现
- 该作业有一定的挑战度，是可选的
- 如果选择这个作业，第3组实验无需验收，也无需提交第3组实验的实验报告

教材及参考书目

- 教材
  - 徐辰，分布式计算系统，高等教育出版社 2022
- 参考书
  - 《设计数据密集型应用》（影印版），Martin Kleppmann著，东南大学出版社，2017
  - 《大数据处理框架Apache Spark设计与实现》许利杰、方亚芬著，电子工业出版社，2020
  - 《大数据计算系统：原理、技术与应用》，王宏志、刘海龙、张立臣、石胜飞编著，机械工业出版社，2023

教学信息

- 教学团队
  - 主讲教师：徐辰
  - 助教：
    - 孙玉书：2022级硕士
    - 刘明熹：2023级硕士
    - 经清源：2020级本科
- 课程信息
  - 钉钉群：课程通知、实验注意事项等
  - 主页：https://dasebigdata.github.io/

2024分布式计算系统

钉钉扫码加入班级

谢谢！ Q&A