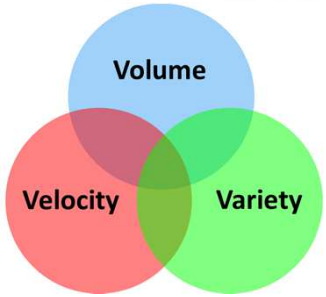
大数据系统与大规模数据分析



陈世敏
（中科院计算所）
孙翼
（国科大计算机学院）

中科院计算技术研究所

陈世敏

博士，中科院计算技术研究所研究员

- 分别于1997年和1999年获清华大学计算机系学士学位和硕士学位，于2005年获卡内基梅隆大学（CMU）计算机科学系博士学位；
- 2005-2013年在美国英特尔研究院、卡内基梅隆大学和惠普研究院任研究员、高级研究员和科研经理；
- 2013年加入中科院计算所，并入选中科院“百人计划”
- 研究方向：数据管理系统、大数据处理、体系结构

课程相关

- 课程时间
  - 每周三晚（9-11节）
  - 课间休息一次：8:10-8:20pm
- 基础知识要求
  - 数据库概论、**程序设计**、数据结构、计算机原理
- 考核方式：闭卷考试+作业
  - 闭卷考试：50%
  - 作业1+作业2+作业3：30%
  - 大作业：20%
  - 课堂表现：+5%

课程安排

周次	内容
9次课	大数据系统 （陈世敏）
5次课	大规模数据分析 （孙翼）
第15周 （6月13日）	期末考试
第16周 （6月20日）	大作业验收，课堂报告

课程安排（大数据系统部分）

周次	内容	
第1周	概论；关系型系统1：关系模型和关系运算	
第2周	关系型系统2：数据库的内部实现	
第3周	关系型系统3：事务处理和数据仓库	
第4周	大数据存储系统1：基础，文件系统，HDFS	作业1
第5周	大数据存储系统2：键值系统	
第6周	大数据运算系统1：MapReduce, 图计算系统	作业2
第8周	大数据存储系统3：图存储，document store	
第9周	大数据运算系统2：图计算系统, MapReduce+SQL	
第10周	大数据运算系统3：内存计算系统	

课程安排（大规模数据分析部分）

周次	内容	
第7周	最邻近搜索和位置敏感（LHS）哈希算法	
第11周	数据空间的维度约化	
第12周	推荐系统	作业3
第13周	流数据采样与估计、流数据过滤与分析	
第14周	大规模数据建模平台应用举例：教育大数据的建模与分析	

上机安排(1)：从第4周开始

- 地点
  - 计算机学院，4层
  - 网络安全教学实验室（447室）：50台
  - 云计算教学实验室（432室）：20台
- 机器：联想PC机M6400t，Windows 7/32bit
  - 环境：每台机器安装了一个虚拟机，运行Ubuntu Linux 14.04.2, Hadoop 2.6.0, HBase 0.98等
  - 作业1-2只需要在单机上构成伪分布环境
  - 大作业可以使用实验室的刀片服务器完成
- 注：可以在自己的计算机上完成作业

上机安排(2)

- 时间
  - 周五上午，8:30-11:50am
  - 周五下午，1:00-4:20pm
- 上机期间助教的职责
  - **管理上机秩序：**上机前找助教签到，分配机器；使用完毕，找助教签出；助教负责监督机房秩序(不得喧哗、打闹等)。
  - **解答机器使用的问题：**包括如何开机、如何登录、如何使用编辑器、如何编译和运行程序
  - **不包括：其它关于作业内容的问题**