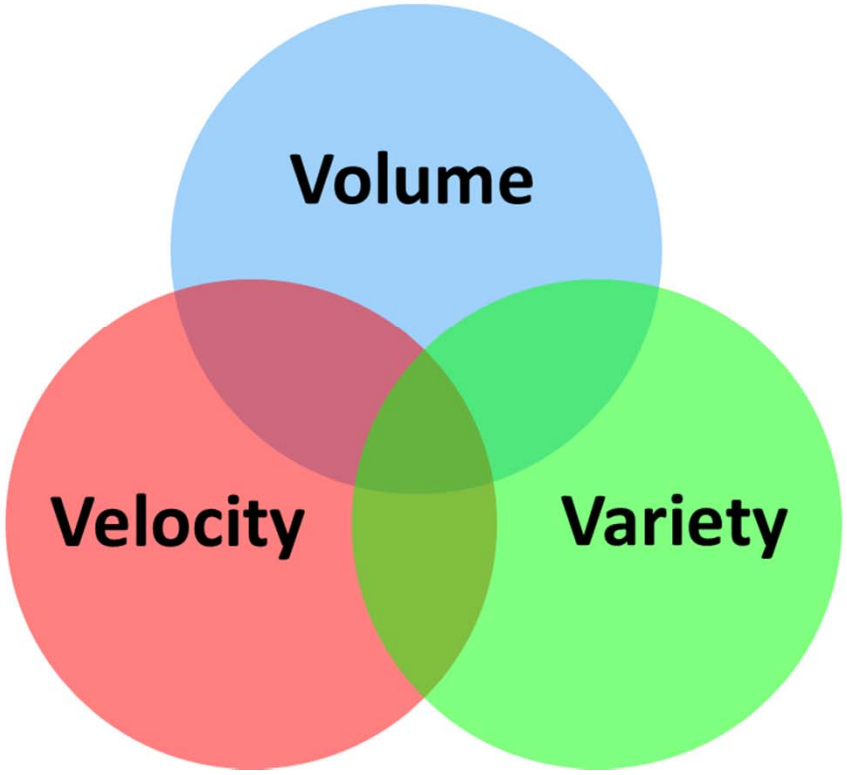
大数据系统与大规模数据分析

大作业



陈世敏
（中科院计算所）

孙翼
（国科大计算机学院）

课程相关

- 成绩分配
  - 闭卷考试：50%
  - 作业1+作业2+作业3：30%
  - 大作业：20%
  - 课堂表现：+3%

作业时间安排

周次	内容	作业
第4周，3/31	大数据存储系统1：基础，文件系统，HDFS	作业1布置
第5周，4/7	大数据存储系统2：键值系统	
第6周，4/14	大数据存储系统3：图存储，document store	
第7周，4/21	大数据运算系统1：MapReduce, 图计算系统	作业1提交 作业2布置
第8周，4/28	大数据运算系统2：图计算系统，MR+SQL	
第9周，5/5=>5/8 (周六上周三的课)	大数据运算系统3：内存计算系统	大作业布置 (系统,6人/组)
第10周，5/12	分布式哈希表，区块链技术中的加密算法	作业2提交
第11周，5/19	最邻近搜索和位置敏感（LSH）算法	作业3
第12周，5/26	奇异值分解与数据空间的维度约化	大作业布置 (分析,3人/组)
第13周，6/2	推荐系统	大作业 仅选1个
第14周，6/9	流数据采样与估计、流数据过滤与分析	
第15周，6/16	期末考试	
第16周，6/23	大作业验收报告（上下午，教1-208）	大作业验收

大作业

- 成绩：占总成绩20%
- 时间
  - 发布：2021/5/8 前
  - Proposal：**2021/5/26 (Wed)，北京时间 11:59pm**
    - 主要目的是确定分组和选题
    - 报告：1页，A4，pdf
    - 内容：选题，成员(选一名组长)，起一个组名，确定每人分工
  - Final Report：**2021/6/23 (Wed)，北京时间 11:59pm**
    - 报告：至少6页，A4，pdf
    - 程序（包括README安装和运行指令）
  - Final Presentation: **2021/6/23 (Wed)**
    - 每组15分钟
    - 目的、文献/现有系统分析、设计、实现、性能/演示等

大数据系统方向：分组和选题

- 自愿组合
  - 大数据系统方向：每组不多于6人
- 每个组
  - 起一个组名（中文或英文），不要太长
  - 选一名组长，组长负责召集组员完成作业
  - 确定成员分工
- 大致确定了作业的选题后，联系助教
  - 希望选不同的题目，允许自选题目
  - **同一个题目，至多3个组可以选**
  - 助教协调，会告知题目是否已经被选了
- 报名：祁琦，qiqi@ict.ac.cn（或者微信群）

机器资源

- 今年华为云将提供免费测试支持
- 具体流程
  - 注册账号
  - 实名认证
  - 报告给祁琦，qiqi@ict.ac.cn
  - 华为云将给账号充值，到课程结束前有效
  - 使用华为云各种服务，例如
    - ECS云虚拟机
    - 华为云数据库服务
- 每组可以有一个免费账号

作业成绩判定：20分

- 5分：新颖性
  - 选择列出的题目自动获得5分
  - 如果是特别具有创新性（在此基础上扩展、自选题目有很强新颖性），可以额外+1分
- 10分：探索深度和实现效果
  - 根据工作量，探索深度，严谨程度，实现程度和效率/演示质量等
- 5分：表达（需要把新颖性和工作深度有效地表达出来）
  - Report的文字是否通顺、易懂
  - PPT讲述是否清楚
- +1分：个人贡献
  - 组长和组员可以推荐具有突出贡献1人，加1分

题目1：数据库对JSON的支持

- 背景
  - 文档数据库以JSON为核心数据类型：例如MongoDB，CouchBase
  - 关系型数据库增加了JSON的支持：例如MySQL，PostgreSQL
- 目标：针对JSON的支持，调研和比较上述4种系统
  - 查找和学习文献（论文、白皮书、技术文档、用户手册）
  - 性能测试和比较
  - 重点内容
    - 系统的基本结构和运行方式？
    - JSON是如何存储的？
    - JSON在查询语句中是如何使用的？
    - 采用现实或合成的JSON数据集（应包括嵌套、数组等结构）来测试4种系统
    - 比较TP类型的点操作（CRUD等）
    - 比较AP类型的数据分析操作（选择、投影、连接、分组聚集等）
    - 考虑两种情况：数据集比内存小，数据集比内存大
  - 形成调研报告

题目2：Spark对JSON的支持

- 背景
  - JSON是一种重要的数据类型，有大量现实数据采用JSON
  - Spark可以通过多种方式来读取和处理JSON数据
    - 读取文本JSON数据
    - 把JSON转化为Parquet文件，读取存储在Parquet中的列式数据
    - 把JSON存储在MongoDB中，读取MongoDB中的JSON数据
    - 近期学术界提出的新技术
- 目标：文献调研，实测使用
  - 查找和学习文献（论文、白皮书、技术文档、用户手册）
    - Spark, Parquet, MongoDB Spark Connector相关文档
    - 阅读近期JSON处理优化的论文至少4篇（SIGMOD, VLDB等）
  - 性能测试和比较
  - 重点内容
    - 对比JSON数据分析的多种方式的工作原理和优化技术
    - 采用现实或合成的JSON数据集（应包括嵌套、数组等结构）来测试Spark
    - 比较Spark+JSON 文本，Spark+Parquet，Spark+Mongo等
    - 比较AP类型的数据分析操作（选择、投影、连接、分组聚集等）
    - 考虑两种情况：单节点、多节点
  - 形成调研报告

题目3：LSM-Tree调研和测试

- 背景
  - LSM-Tree从1996年提出，至今被广泛应用
  - 文献和实际系统中出现了多种结构方式和Compaction形式
- 目标：文献调研和性能测试
  - 查找和学习文献（论文、白皮书、技术文档）
  - 性能测试和比较
  - 重点内容
    - 加深学习关于LSM-Tree的课程内容，阅读相关论文
      - SIGMOD, VLDB, ICDE, SOSP, OSDI等，阅读不少于6篇论文（侧重近期）
      - 结构方式，Compaction形式
    - 调研典型系统RocksDB和Cassandra
      - 结构方式，Compaction形式，参数设置
    - 产生数据测试RocksDB和Cassandra
      - 分析Compaction对于性能的影响
      - 考虑数据的不同情况：例如，append-only，大量修改或删除，等情况
  - 形成调研报告

题目4：云TP数据库调研

- 产业界推出了多种云TP数据库解决方案
  a) Amazon Aurora
  b) 阿里 PolarDB
  c) 腾讯TDSQL-C
  d) PingCap TiDB
- 目标：文献调研和测试
  - 查找和学习文献（论文、白皮书、技术文档、用户手册）
    - 系统有相关论文的必须阅读
  - 实测华为云数据库性能
  - 重点内容
    - 调研上述4种TP数据库的系统结构、工作原理
    - 从课程内容出发，比较上述4种系统
    - 实测华为云提供的数据库MySQL服务的性能
      - 采用TP类型的测试集：例如YCSB，TPCC
      - 与ECS自己部署的MySQL进行对比
  - 形成调研报告

题目5：云AP数据库调研

- 产业界推出了多种云AP数据库解决方案
  a) Amazon Redshift
  b) 阿里 AnalyticDB
  c) 腾讯TDSQL-A
  d) Snowflake
- 目标：文献调研和测试
  - 查找和学习文献（论文、白皮书、技术文档、用户手册）
    - 系统有相关论文的必须阅读
  - 实测华为云数据库性能
  - 重点内容
    - 调研上述4种AP数据库的系统结构、工作原理
    - 从课程内容出发，比较上述4种系统
    - 实测华为云提供的数据库PostgreSQL服务/GuassDB(OpenGauss)的性能
      - 采用AP类型的测试集：例如TPCH
      - 与ECS自己部署的PostgreSQL进行对比
  - 形成调研报告

题目6：Serverless Computing调研

- Serverless Computing
  - 普通的云平台用户分配虚拟机，配置所需的服务，然后运行应用
  - Serverless computing的用户只需要提交应用，云平台负责分配运行资源
- Serverless Runtime: function as a service平台
  - Amazon AWS Lambda，Google Cloud Functions等
  - 华为云FunctionGraph
- 目标：文献学习调研Serverless Computing，并试用
  - 查找和学习文献（论文、白皮书、技术文档、用户手册）
  - 实测试用
  - 重点内容
    - 阅读近期相关论文不少于6篇（OSDI, SOCC, SIGMOD等）
    - 学习Serverless computing：系统结构、工作原理、编程界面、发展趋势
    - 试用华为云FunctionGraph
      - 选用研读论文中的实验用例或者华为云提供的案例
      - 测试FunctionGraph的性能、弹性可扩展性等
  - 形成调研报告

题目7：Approximate Query Processing

- 背景
  - 在大量数据上的分析运算时间较长，无法满足交互要求
  - 一种方法是采用一个小的样本集合，然后在小样本上运行
  - Approximate query processing
- 思路
  a) 采用Spark提供的sample()和sampleByKey()接口
  b) 事先生成sample，然后在生成的sample上运行
- 目标：文献调研+实验分析
  - 调研Approximate Query Processing相关文献
    - 综述、论文等（SIGMOD, VLDB, ICDE等），至少6篇论文
  - 采用公开或合成的数据集（例如TPCH）
  - 实验测试性能与准确性
    - Join对结果有什么影响？如何有效的支持Join？
  - 能否把上述两种思路结合起来？

题目8：相似查询分析

- 背景
  - 数据库SQL：要求给出准确的查询条件，返回准确计算的结果
  - 文本搜索：根据关键字，按照某种规则（例如TFIDF，PageRank）排序返回结果
  - 在有些情况下，这两类操作可能都不很完美
- 举例1：医院数据库包括了患者的多种信息，例如

给定病人张三，
医生想找到与张三
相似的其他病例？

给定一个病例描述，
医生想找到相似的
其他病例？

题目8：相似查询分析

- 举例2：电影数据库包括了电影的信息

给定一部电影，
找到类似的电影？

给定电影描述，
找到类似的电影？

电影信息
导演、演
员信息等

IMDb movies.csv (45.71 MB)

Detail | Compact | Column | 10 of 22 columns

imdb_title_id	title	original_title	year	date_publi...	genre	duration
tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45
tt0000574	The Story of the Kelly Gang	The Story of the Kelly Gang	1906	1906-12-26	Biography, Crime, Drama	70
tt0001892	Den sorte drøm	Den sorte drøm	1911	1911-08-19	Drama	53
tt0002101	Cleopatra	Cleopatra	1912	1912-11-13	Drama, History	100
tt0002130	L'Inferno	L'Inferno	1911	1911-03-06	Adventure, Drama, Fantasy	68
tt0002199	From the Manger to the Cross; or, Jesus of Nazareth	From the Manger to the Cross; or, Jesus of Nazareth	1912	1913	Biography, Drama	60
tt0002423	Madame DuBarry	Madame DuBarry	1919	1919-11-26	Biography, Drama, Romance	85
tt0002445	Quo Vadis?	Quo Vadis?	1913	1913-03-01	Drama, History	120
tt0002452	Independenta Romaniei	Independenta Romaniei	1912	1912-09-01	History, War	120
tt0002461	Richard III	Richard III	1912	1912-10-15	Drama	55

类似“流浪星球”的电影

题目8：相似查询分析

- 思路1：SQL only
  - 把给定记录/记录描述转换为过滤条件
    - 流浪星球➔提取属性，例如科幻片，吴京
    - 把这些信息组成SQL语句
  - 使用SQL完成过滤运算
- 思路2：Keyword search
  - 在关系表上建立文本倒排索引
  - 把给定记录/记录描述转换为关键字搜索条件
  - 通过关键字搜索来找到对应的记录
- 思路3：graph search
  - 在关系表基础上建立一个属性图
  - 把给定记录/记录描述转换为子图匹配+图查询
  - 通过图搜索来找到对应的记录

题目8：相似查询分析

- 目标：文献调研+方案分析
  - 对相关领域进行文献调研
  - 采用公开数据集（例如：https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv）
  - 实现数据集的关系表、倒排索引、图数据库结构
    - 选用开源系统实现
  - 采用思路1、2、3来支持相似查询操作
    - 简单方法：对一个给定的查询，手工写出对应的查询语句
    - 高级方法：对一个给定的查询，自动生成查询语句
      - 可以完成一部分：比如，对于某个思路来完成，或者限定自动生成的能力范围
    - 比较三种思路的效果
  - 可选：其他更好的方案？（比如AI模型？）
  - 通过上述调研与实验分析，对相似查询的支持进行比较和分析

大数据系统：大作业题目

- 题目1：数据库对JSON的支持
- 题目2：Spark对JSON的支持
- 题目3：LSM-Tree调研和测试
- 题目4：云TP数据库调研
- 题目5：云AP数据库调研
- 题目6：Serverless Computing调研
- 题目7：Approximate Query Processing
- 题目8：相似查询分析

数据分析：大作业题目

- 孙翼老师会进一步发布选题
- 论文学习理解
- 每组3人，论文不重复
- 大致10组

大作业

- 成绩：占总成绩20%
- 时间
  - 发布：2021/5/8 前
  - Proposal：**2021/5/26 (Wed)，北京时间 11:59pm**
    - 主要目的是确定分组和选题
    - 报告：1页，A4，pdf
    - 内容：选题，成员(选一名组长)，起一个组名，确定每人分工
  - Final Report：**2021/6/23 (Wed)，北京时间 11:59pm**
    - 报告：至少6页，A4，pdf
    - 程序（包括README安装和运行指令）
  - Final Presentation: **2021/6/23 (Wed)**
    - 每组15分钟
    - 目的、文献/现有系统分析、设计、实现、性能/演示等

报名：祁琦，qiqi@ict.ac.cn （或者微信群）