NoSQL数据库技术项目作业

讲师：刘奎恩 博士

项目作业

- 宏观行业数据统计
- 新浪微博数据存取
- 基于Web的数据存取

- 选修题
  - 在Redis或MongoDB的社区提交一个PR
    - 即，修改哪怕一行代码，并申请并入master分支

问题描述——宏观行业数据统计

- 宏观行业指标数据
  - 宏观行业经济指标是体现经济情况的一种方式，主要指标包括国民生产总值、通货膨胀与紧缩、投资指标、消费、金融、财政指标等，对于宏观经济调控起着重要的分析和参考作用。
  - 常用缩写
    - CPI：消费者物价指数(Consumer Price Index)。
    - PPI：生产者物价指数（Producer Price Indexes）或产品价格指数。
    - PMI：采购经理指数（Purchase Management Index）。
  - 问题：
    - 经济是否处于通膨？
    - 经济是否过热？
    - 失业是否在加剧？

数据描述

- 数据
  - 中国宏观主要经济指标，如GDP、PMI、CPI、PPI、景气指数、工业、国内贸易、居民收支、外商投资、进出口、房地产开发、固定资产投资，利率及财政相关数据。
  - 美国宏观数据，如GDP, 价格指数，就业失业及货币供应量。详见子目录及子目录描述。
  - 数据格式：Excel文件

```
kain:宏观行业 kliu$ ls -lh
total 288
-rw-r--r--@ 1 kliu  staff    48K May 21  2015 CPI.csv
-rw-r--r--@ 1 kliu  staff   4.2K May 21  2015 GDP.csv
-rw-r--r--@ 1 kliu  staff    46K May 21  2015 PMI.csv
-rw-r--r--@ 1 kliu  staff   2.6K May 21  2015 美国GDP.csv
-rw-r--r--@ 1 kliu  staff    30K May 21  2015 美国就业与失业.csv
```

数据格式(GDP)

数据	指标名称	数据格式（举例）
GDP	指标代码,	G010000007
	指标全称,	季_美国_GDP价格指数_季调
	数据发布时间	2015-01-30 08:30:00
	数据日期	2014-12-31
	数据值,	108.68
	单位,	2009年=100,
	地区,	美国,
	数据来源	美国国家经济分析局
	数据更新时间	2015-04-10 14:21:54

数据格式(CPI)

数据	指标名称	数据格式（举例）
CPI	指标代码,	M030000003
	指标全称,	月_居民消费价格指数(CPI)_同比
	数据发布时间	2015-01-09 09:30:00
	数据日期	2014-12-31
	数据值,	1.51
	单位,	%
	地区,	中国
	数据来源	国家统计局
	数据更新时间	2015-01-09 18:41:51

任务描述

- 利用所学到的NoSQL技术，加载并进行数据处理，包括：
  - 经济是否处于通膨？
    - 当CPI>3%时，我们称为INFLATION，就是通货膨胀；
    - CPI>5%的增幅时，称为SERIOUS INFLATION，就是严重通货膨胀。
  - 经济是否过热？
    - CPI/PPI上涨（过高），一般是经济偏热（过热）的表面特征
    - PMI与上月进行比较，大于50%，表示经济上升，反之则趋向下降。
  - 失业是否在加剧？
    - 统计每个季度的农业和非农业失业率以及其变化
- 要求：
  - 在MongoDB上实现

作业提交形式

- 报告（描述完成该任务的详细过程），应该包括以下四个部分；
  - 问题描述；
  - 数据加载代码；
  - 数据统计算法；
  - 分析结果及对结果的解释；
- 源代码；

问题描述——微博数据处理

- 宏观行业指标数据
  - 宏观行业经济指标是体现经济情况的一种方式，主要指标包括国民生产总值、通货膨胀与紧缩、投资指标、消费、金融、财政指标等，对于宏观经济调控起着重要的分析和参考作用。
  - 常用缩写
    - CPI：消费者物价指数(Consumer Price Index)。
    - PPI：生产者物价指数（Producer Price Indexes）或产品价格指数。
    - PMI：采购经理指数（Purchase Management Index）。
  - 问题：
    - 经济是否处于通膨？
    - 经济是否过热？
    - 失业是否在加剧？

数据描述

- 新浪微博用户信息
  - 数据格式：用户uid，用户昵称，用户姓名，用户所在地，用户主页url，用户性别，用户粉丝数，用户关注数，用户微博数，用户收藏数，用户创建时间；
- 新浪微博信息
  - 数据格式：微博mid，发布时间，微博内容，微博来源，微博转发数，微博评论数，微博被赞数，发表用户uid，微博所属主题。
  - 12个主题包括魅族，小米，火箭队，林书豪，恒大，韩剧，雾霾，房价，同桌的你，公务员，贪官，转基因。
- 新浪微博用户好友关系
  - 每条记录由suid和tuid两个字段组成，表示suid关注tuid，由于新浪微博的限制，每个用户最多只能获取到200个关注人的信息，故好友关系不是很全；
- 新浪微博转发关系
  - 每条记录由smid和tmid两个字段组成，表示smid微博转发tmid微博。

数据描述

- 原始数据格式
  - 数据文件： 数据文件为sql脚本，需要转换为nosql格式

- 数据规模
  - 63641个新浪微博用户信息
  - 84168条微博信息
    - 在2014年05月3-11日采集
    - 关于12个主题
  - 1391718条用户好友关系
  - 27759条微博转发关系

数据描述

数据	指标名称	数据格式（举例）
用户信息	uid	'1860096194'
	Screen_name	'耿琪伟'
	name	'耿琪伟',
	province	32
	city	1
	location	'江苏 南京'
	url	'http://blog.sina.com.cn/caaaaaasal'
	gender	'm'
	followersnum	149971
	friendsnum	182
	statusesnum	7387
	favouritesnum	4
	created_at	'2010-10-22 17:42:00

数据描述

数据	指标名称	数据格式（举例）
微博信息	mid	'3706920600804587'
	date	'2014-05-05 17:30:05'
	text	'继电视之后，路由器将成为下一个被玩烂的所谓智能产品。。。乐视也要做路由器，目测又要和小米路由大战三百回合。。。',
	source	'微博 weibo.com'
	repostsnum	5
	commentsnum	9
	attitudesnum	3
	uid	‘1860096194'
	topic	'小米'

数据描述

- 用户关系
  - 耿琪伟 → 经纬张颖

数据	指标名称	数据格式（举例）
用户关系	suid	'1860096194'
	tuid	‘1042026447'

- 微博关系

数据	指标名称	数据格式（举例）
用户关系	suid	'3706920600804587'
	tuid	‘3705035298678486'

任务描述

- 利用所学到的NoSQL技术，加载并进行数据处理，包括：
  - 找出发帖子最多的5个用户？
  - 找出转发次数最多的20个帖子？
  - 找出最热门的3个话题？
  - 找出具有最多相同好友的5对用户？
  - 找出转发最多相同帖子的5对用户？

- 要求：
  - 在Redis上实现

作业提交形式

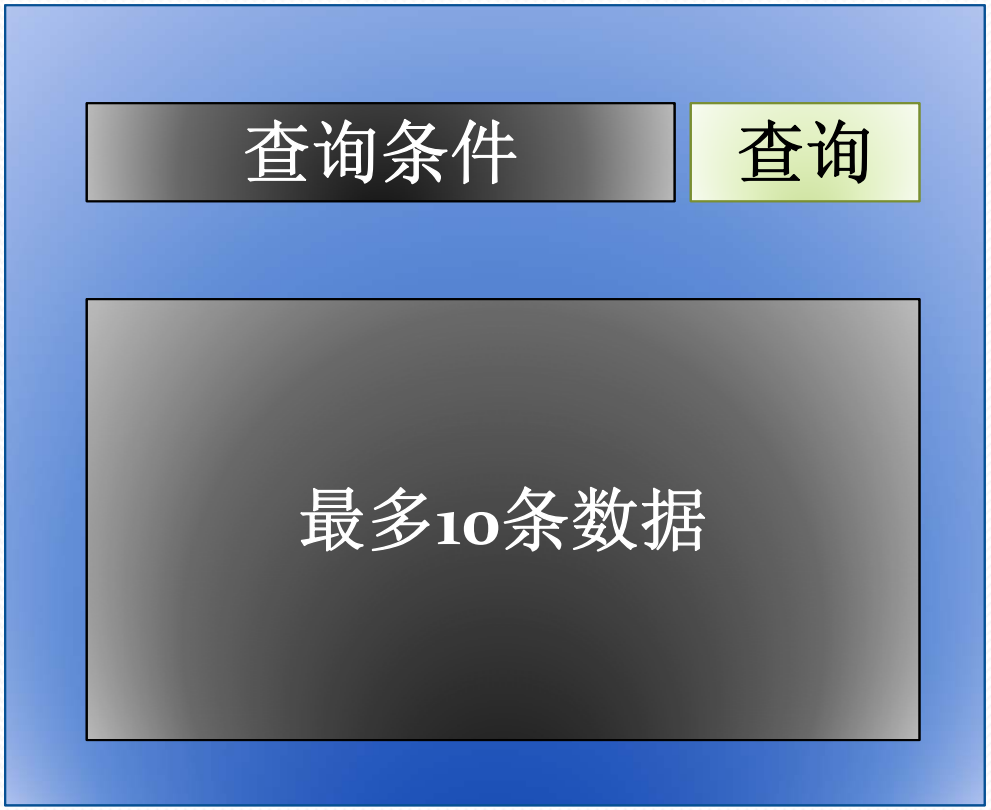
- 报告（描述完成该任务的详细过程），应该包括以下四个部分；
  - 问题描述；
  - 数据加载代码；
  - 数据统计算法；
  - 分析结果及对结果的解释；
- 源代码；

问题描述——基于Web的数据存取

- NoSQL的兴起伴随着互联网的高速发展，以NoSQL作为数据后台可以快速搭建Web应用。
- 已有数据源
  - 宏观行业数据统计
  - 新浪微博数据存取
- 设计一个架构，后台用NoSQL进行数据存储，前台可以通过Web进行数据查询。

任务描述

- 利用所学到的NoSQL技术，加载并进行Web查询，包括：
  - 一个Web页面，有2部分
    - 第一部分
      - 数据查询条件输入
    - 第二部分
      - 数据查询结果显示
- 要求：
  - 利用已有数据：
    - 宏观行业数据统计
    - 新浪微博数据存取



作业提交形式

- 报告（描述完成该任务的详细过程），应该包括以下四个部分；
  - 问题描述；
  - 数据加载代码；
  - 数据统计算法；
  - 分析结果及对结果的解释；
- 源代码；

谢谢