SupStat

R 语言

统计分析、绘图、数据挖掘

hadoop

# R语言为Hadoop注入统计血脉 - 张丹

2013.11.23

北京数博思达信息科技有限公司

# 目录

- R语言介绍
- Hadoop介绍
- 为什么要让Hadoop结合R语言？
- 如何让Hadoop结合R语言？
- R和Hadoop结合的案例
- 演示程序
- 展望未来

# R语言介绍：起源

- R语言，一种自由软件编程语言，主要用于统计分析、绘图、数据挖掘。
- R是由来自新西兰奥克兰大学的Ross Ihaka和Robert Gentleman开发，现在由“R开发核心团队”负责开发。
- R是基于S语言的一个GNU计划项目，所以也可以当作S语言的一种实现。
- R的语法是来自Scheme。

# R语言介绍：跨平台，许可证

- R的源代码可自由下载使用，GNU通用公共许可证。
- R可在多种平台下运行，包括UNIX,Linux,Windows和MacOS。R主要是以命令行操作为主，同时支持GUI的图形用户界面。

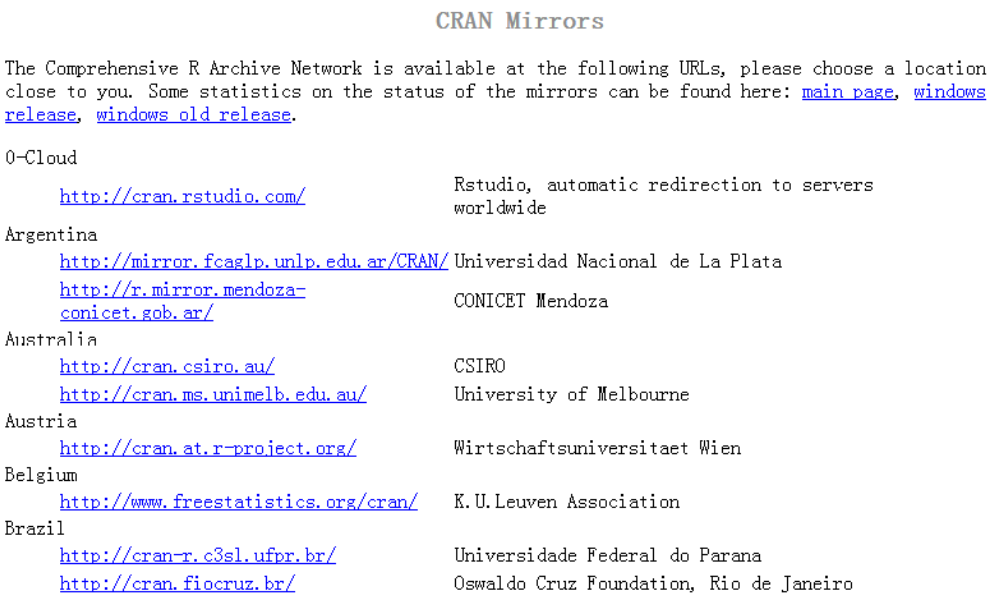
- GNU许可证：商业软件可以使用，但不能修改LGPL协议的代码。
- http://blog.fens.me/it-license/

# R语言介绍： R的数字基因

- R内建多种统计学及数字分析功能。因为S的血缘，R比其他统计学或数学专用的编程语言有更强的面向对象功能。
- R的另一强项是绘图功能，用R做数据可视化输出，达到专业级水平。
- 用R作矩阵计算，其分析速度可媲美GNU Octave，甚至商业软件MATLAB。
  - Blas调优后，R可以在底层实现矩阵的并行计算

# R语言介绍：代码库

- CRAN为Comprehensive R Archive Network的简称。它除了收藏了R的执行档下载版、源代码和说明文件，也收录了各种用户撰写的软件包。全球有超过一百个CRAN镜像站，上万个第三方的软件包。
- http://cran.r-project.org/mirrors.html



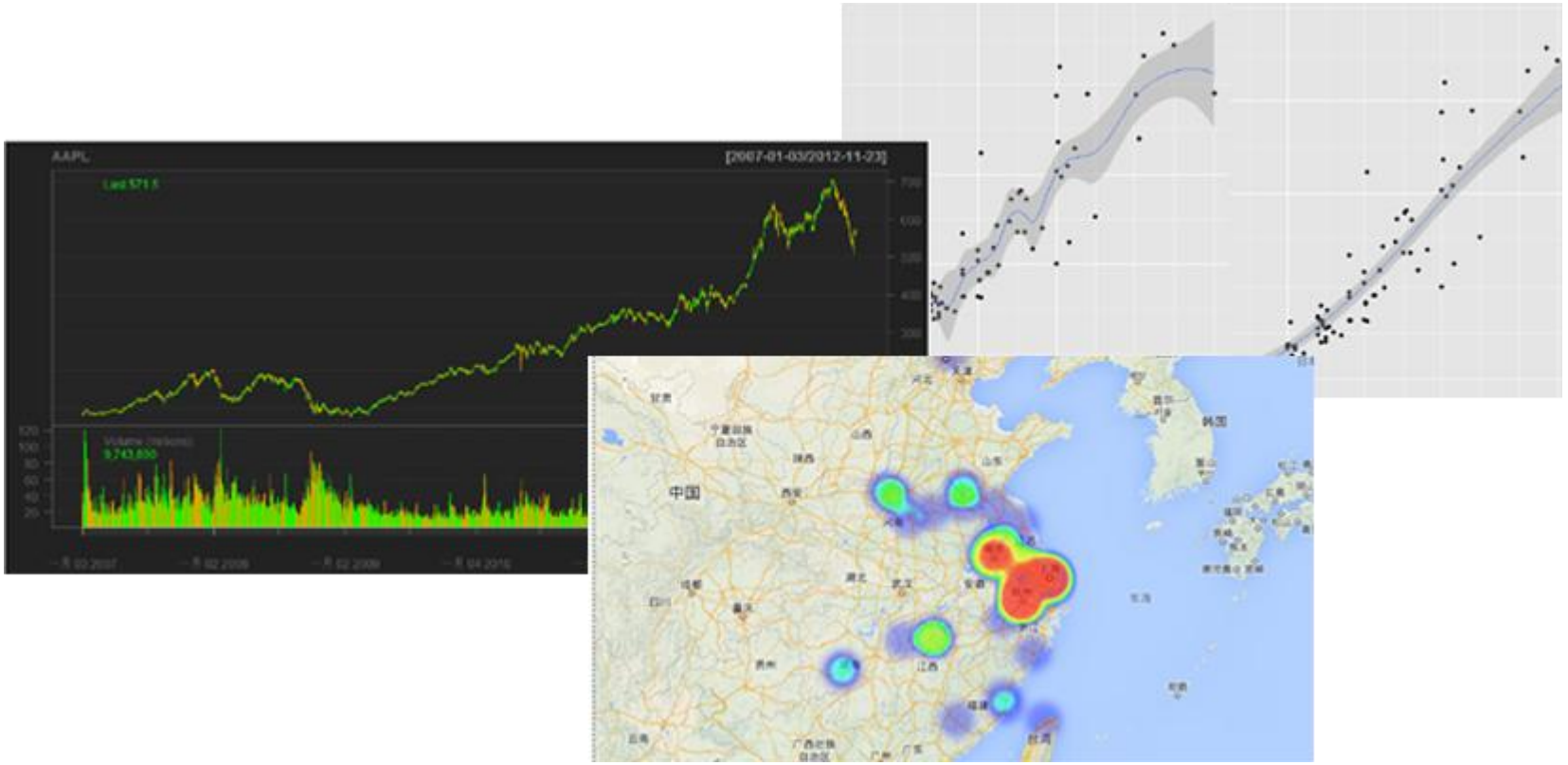
CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

| | |
|---|---|
| 0-Cloud | |
| http://cran.rstudio.com/ | Rstudio, automatic redirection to servers worldwide |
| Argentina | |
| http://mirror.fcaglp.unlp.edu.ar/CRAN/ | Universidad Nacional de La Plata |
| http://r.mirror.mendoza-conicet.gob.ar/ | CONICET Mendoza |
| Australia | |
| http://cran.csiro.au/ | CSIRO |
| http://cran.ms.unimelb.edu.au/ | University of Melbourne |
| Austria | |
| http://cran.at.r-project.org/ | Wirtschaftsuniversitaet Wien |
| Belgium | |
| http://www.freestatistics.org/cran/ | K.U.Leuven Association |
| Brazil | |
| http://cran-r.c3sl.ufpr.br/ | Universidade Federal do Parana |
| http://cran.fiocruz.br/ | Oswaldo Cruz Foundation, Rio de Janeiro |

# R语言介绍：R的行业应用

- 统计分析，应用数学，计量经济，金融分析，财经分析，人文科学，数据挖掘，人工智能，生物信息学，生物制药，全球地理科学，数据可视化。



# R语言介绍：商业竞争对手

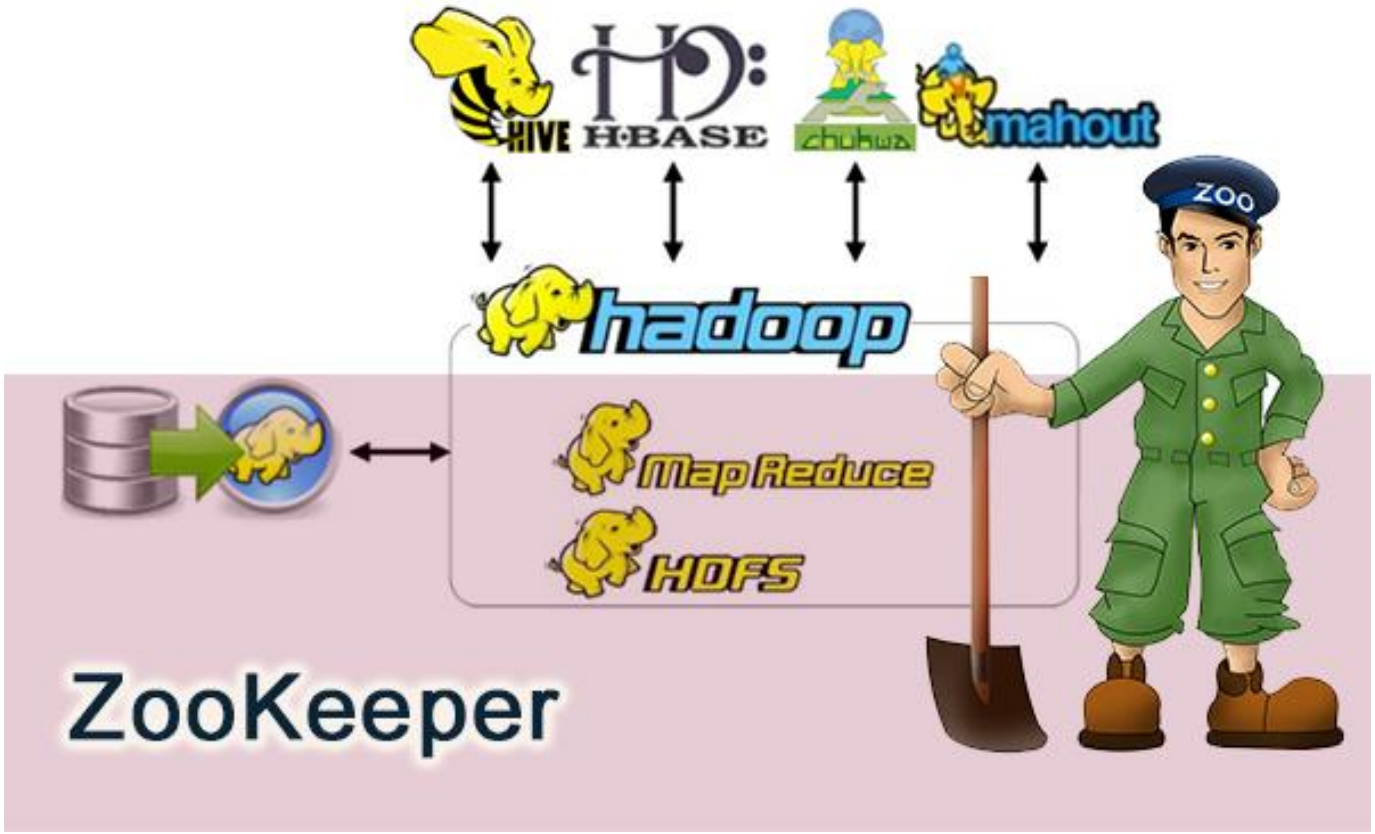
- SAS:(Statistical Analysis System),是SAS公司推出的一款用于数据分析和和决策支持的大型集成式模块化软件系统。
- SPSS:（Statistical Product and Service Solutions）是IBM公司推出的一系列用于统计学分析运算、数据挖掘、预测分析和决策支持任务的软件产品及相关服务的总称。
- Matlab: (MATrix LABoratory)，是MathWorks公司出品的一款商业数学软件。MATLAB是一种用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境。

# Hadoop介绍

- Hadoop是一个分布式系统基础架构，由Apache基金会开发。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力高速运算和存储。Hadoop实现了一个分布式文件系统（Hadoop Distributed File System），简称HDFS。HDFS有着高容错性的特点，并且设计用来部署在低廉的（low-cost）硬件上。而且它提供高传输率（high throughput）来访问应用程序的数据，适合那些有着超大数据集（large data set）的应用程序。HDFS放宽了（relax）POSIX的要求（requirements）这样可以流的形式访问（streaming access）文件系统中的数据。

# Hadoop介绍：家族成员

- Hive, HBase, Zookeeper, Avro, Pig, Ambari, Sqoop, Mahout, Chukwa



# Hadoop介绍：家族成员

- Hive: 是基于Hadoop的一个数据仓库工具。
- Pig: 是一个基于Hadoop的大规模数据分析工具。
- HBase: 是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统。
- Sqoop: 是一个用来将Hadoop和关系型数据库中的数据相互转移的工具。
- Zookeeper:是一个为分布式应用所设计的分布的、开源的协作系统。
- Mahout:是基于Hadoop的机器学习和数据挖掘的一个分布式框架。
- Avro: 是一个数据序列化系统，设计用于支持数据密集型，大批量数据交换的应用。
- Ambari: 是一种基于Web的工具，支持Hadoop集群的供应、管理和监控。
- Chukwa: 是一个开源的用于监控大型分布式系统的数据收集系统。

http://blog.fens.me/hadoop-family-roadmap/

# Hadoop介绍：家族成员

- 自2006年，Hadoop以MapReduce和HDFS独立发展开始，到今年2013年不过7年时间，Hadoop的家族已经孵化出多个Apache的顶级项目。特别是最近1-2年，发展速度越来越快，又融入了很多新技术(YARN, Hcatalog, Oozie, Cassandra)，都有点让我们都学不过来了。

# 为什么要让Hadoop结合R语言？

- R语言介绍和Hadoop，在各自领域的都占有重要地位。很多开发人员在计算机的角度，都会提出下面2个问题。

- 问题1: Hadoop的家族如此之强大，为什么还要结合R语言？
- 问题2: Mahout同样可以做数据挖掘和机器学习，和R语言的区别是什么？

# 问题1: Hadoop的家族如此之强大，为什么还要结合R语言？

- Hadoop家族的强大之处，在于对大数据的处理，让原来对TB,PB数据量计算 的不可能，成为了可能。
- R语言的强大之处，在于统计分析，在没有Hadoop之前，我们对于大数据的处理，要取样本，假设检验，做回归。长久以来R语言都是统计学家专属的工具。
- 从上面两点，我们可以看出，hadoop重点是**全量数据分析**，而R语言重点是**样本数据分析**。 两种技术放在一起，刚好是最长补短！

- 模拟场景：**对1PB的新闻网站访问日志做分析，预测未来流量变化**
- 如何实现这样的系统呢？

# 问题1: Hadoop的家族如此之强大，为什么还要结合R语言？

1. 用R语言，通过分析少量数据，对业务目标建模(回归分析)，并定义指标。
2. 用Hadoop从海量日志数据中，提取指标数据
3. 用R语言建模，对指标数据进行测试，验证和调优
4. 用Hadoop分步式算法，重写R语言的模型
5. 部署上线

# 问题1: Hadoop的家族如此之强大，为什么还要结合R语言？

- 这个场景中，R和Hadoop分别都起着非常重要的作用。
  - 以计算机开发人员的思路，所有有事情都用Hadoop去做，会没有数据建模和证明的过程，“预测的结果”一定是有问题的。
  - 以统计人员的思路，所有的事情都用R去做，以抽样方式对样本数据分析，得到的“预测的结果”也一定是有问题的。

- 所以让二者结合，是产界业的必然的导向，也是产界业和学术界的交集，同时也为交叉学科的人才提供了无限广阔的想象空间。

# 问题2: Mahout同样可以做数据挖掘和机器学习，和R语言的区别是什么？

- Mahout是基于Hadoop的数据挖掘和机器学习的算法框架，Mahout的重点同样是解决大数据的计算的问题。
- Mahout目前已支持的算法包括，协同过滤，推荐算法，聚类算法，分类算法，LDA, 朴素Bayes，随机森林。上面的算法中，大部分都是距离的算法，可以通过矩阵分解后，充分利用MapReduce的并行计算框架，高效地完成计算任务。
- Mahout的空白点：
  - 还有很多的数据挖掘算法，很难实现MapReduce的并行化。
  - Mahout的现有模型，都是通用模型，直接用到的项目中，计算结果只会比随机结果好一点点，算法优化的难度很大。
  - Mahout二次开发，要求有深厚的JAVA和Hadoop的技术基础，并最好兼有 “线性代数” ，“概率统计” ，“算法导论” 等的基础知识。所以想玩转Mahout真的不是一件容易的事情。

# 问题2: Mahout同样可以做数据挖掘和机器学习，和R语言的区别是什么？

- R语言同样提供了Mahout支持的绝大多数算法(除专有算法)，并且还支持大量的Mahout不支持的算法，算法的更新速度比Mahout快N倍。并且开发简单，参数配置灵活，对小型数据集运算速度非常快。

- 虽然，Mahout同样可以做数据挖掘和机器学习，但是和R语言的擅长领域并不重合。集百家之长，在适合的领域选择合适的技术，才能真正地“保质保量”做软件做分析。

# 如何让Hadoop结合R语言？

- **RHadoop**

  RHadoop是一款Hadoop和R语言的结合的产品，由RevolutionAnalytics公司开发，并将代码开源到github社区上面。RHadoop包含三个R包 (rmr，rhdfs，rhbase)，分别是对应Hadoop系统架构中的，MapReduce, HDFS, HBase 三个部分。

- 参考文章:

  RHadoop实践系列之二 RHadoop安装与使用

  RHadoop实践系列之四 rhbase安装与使用

REVOLUTION ANALYTICS

THINKBIG ANALYTICS

# 如何让Hadoop结合R语言？

- **RHive**

  RHive是一款通过R语言直接访问Hive的工具包，是由NexR一个韩国公司研发的。

- 参考文章:

  R利剑NoSQL系列文章 之 Hive

  用RHive从历史数据中提取逆回购信息

# 如何让Hadoop结合R语言？

- **重写Mahout**

  用R语言重写Mahout，也是一种结合的思路，我也做过相关的尝试。

- 参考文章:

  用R解析Mahout用户推荐协同过滤算法(UserCF)

# 如何让Hadoop结合R语言？

- **Hadoop调用R**
- 上面说的都是R如何调用Hadoop，当然我们也可以反相操作，打通JAVA和R的连接通道，让Hadoop调用R的函数。但是，这部分还没有商家做出成形的产品。

- 我写了2个例子，大家可以自己尝试着结合，做出不一样的应用来。
- 参考文章:

  Rserve与Java的跨平台通信

  解惑rJava R与Java的高速通道

# R和Hadoop在实际中的案例

- R和Hadoop的结合，技术门槛还是有点高的。对于一个人来说，不仅要掌握Linux, Java, Hadoop, R的技术，还要具备 软件开发，算法，概率统计，线性代数，数据可视化，行业背景 的一些基本素质。

- 在公司部署这套环境，同样需要多个部门，多种人才的的配合。Hadoop运维，Hadoop算法研发，R语言建模，R语言MapReduce化，软件开发，测试等等。。。

# R和Hadoop在实际中的案例

- 这样的案例并不太多。
- 我做过一些尝试和努力，已经整理成文章的有3个项目。文章只是介绍了如何使用Rhadoop的技术，不是完整的系统。

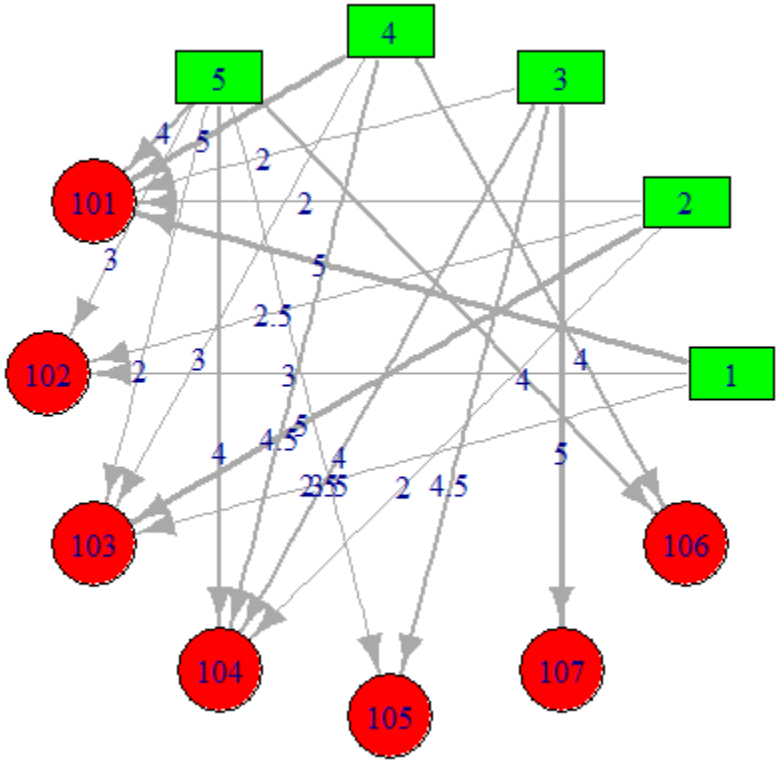
- 参考文章：
- RHadoop实践系列之三 R实现MapReduce的协同过滤算法
  RHadoop实验 – 统计邮箱出现次数
  用RHive从历史数据中提取逆回购信息

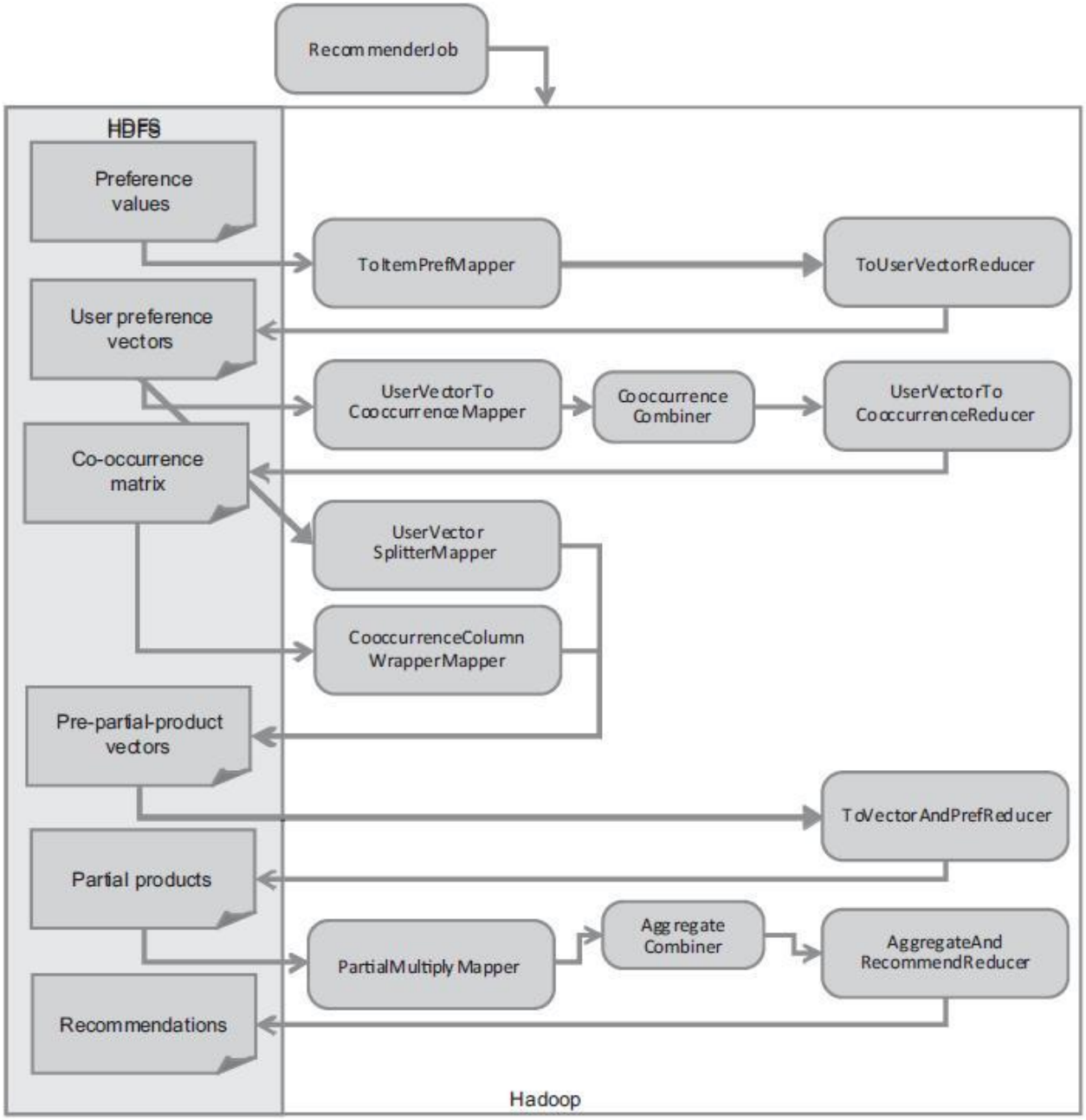
# 演示程序

- RHadoop基础程序
- 分步式协同过滤ItemCF算法介绍
- R本地程序实现
- RHadoop实现
- Java Hadoop MapReduce实现
- Mahout 实现

# 测试数据

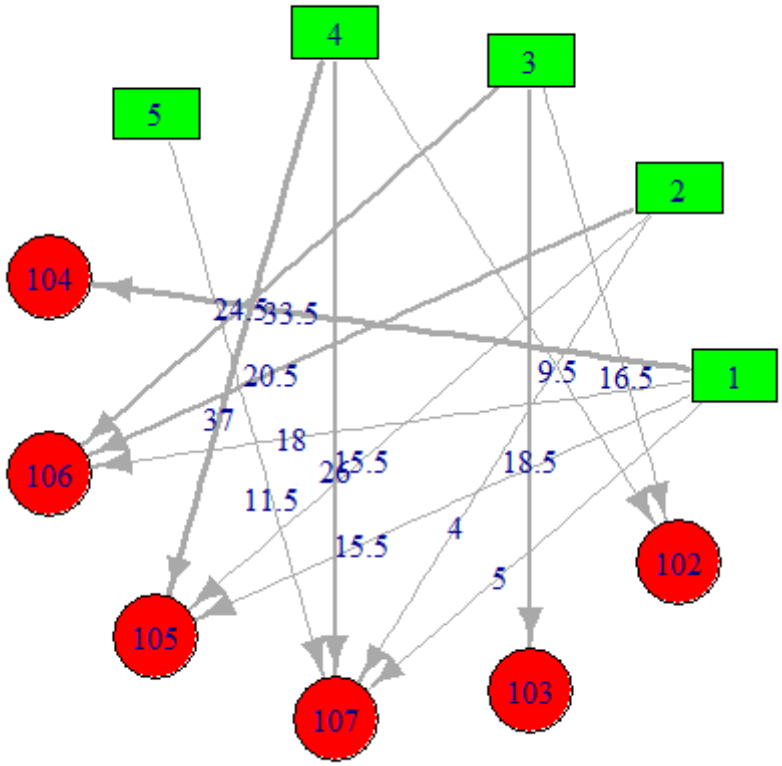
1,101,5.0
1,102,3.0
1,103,2.5
2,101,2.0
2,102,2.5
2,103,5.0
2,104,2.0
3,101,2.0
3,104,4.0
3,105,4.5
3,107,5.0
4,101,5.0
4,103,3.0
4,104,4.5
4,106,4.0
5,101,4.0
5,102,3.0
5,103,2.0
5,104,4.0
5,105,3.5
5,106,4.0




RecommenderJob
HDFS
Preference values
ToItemPrefMapper
ToUserVectorReducer
User preference vectors
UserVectorTo CooccurrenceMapper
Cooccurrence Combiner
UserVectorTo CooccurrenceReducer
Co-occurrence matrix
UserVector SplitterMapper
CooccurrenceColumn WrapperMapper
Pre-partial-product vectors
ToVectorAndPrefReducer
Partial products
PartialMultiply Mapper
Aggregate Combiner
AggregateAnd RecommendReducer
Recommendations
Hadoop

# 推荐结果

1,104,33.5
1,106,18
1,105,15.5
1,107,5
2,106,20.5
2,105,15.5
2,107,4
3,103,24.5
3,102,18.5
3,106,16.5
4,102,37
4,105,26
4,107,9.5
5,107,11.5



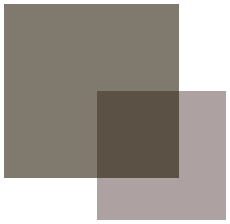
# 展望未来

- 对于R和Hadoop的结合，在近几年，肯定会生成爆发式的增长的。但由于跨学科会造成技术壁垒，人才会远远跟不上市场的需求。

- 所以，肯定会有更多的大数据工具，被发明！
- 机会就在我们的手中，也许明天你的创新，就是我们追逐的方向！！

# 联系作者

- 张丹, 编程爱好者(Java,R,PHP,Javascript)
- Weibo:  @Conan_Z
- Blog：**http://blog.fens.me**
- Email: bsspirit@gmail.com



SupStat™

# Thanks

## FAQ时间