

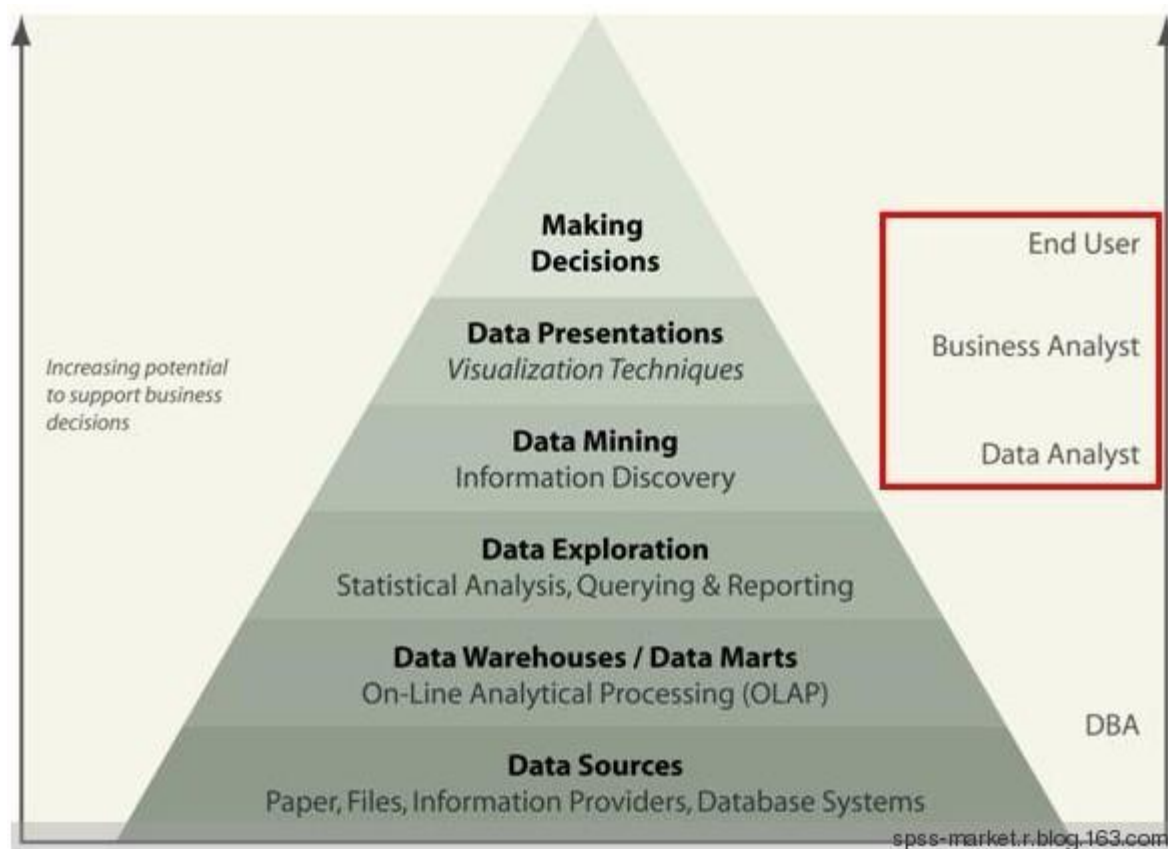


Hadoop应用开发实战案例 第1周

【声明】 本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

<http://edu.dataguru.cn>



- 业务人员
- ETL工程师
- 数据仓库工程师
- 数据分析师
- 数据展现设计师
- IT支持人员：运维，程序员，生产线数据管理员

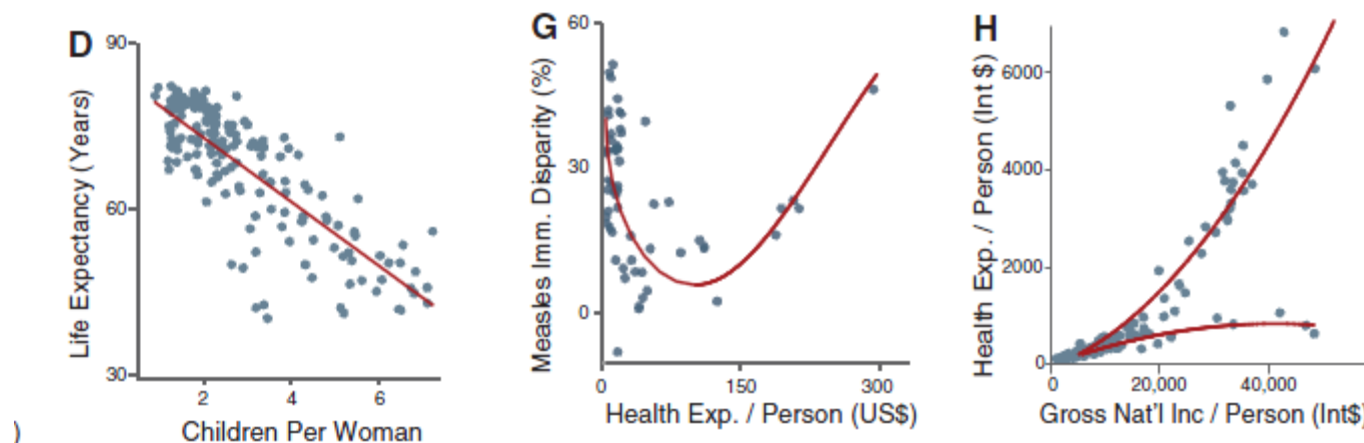


- 使用统计方法，有目的地对收集到的数据进行分析处理，并且解读分析结果

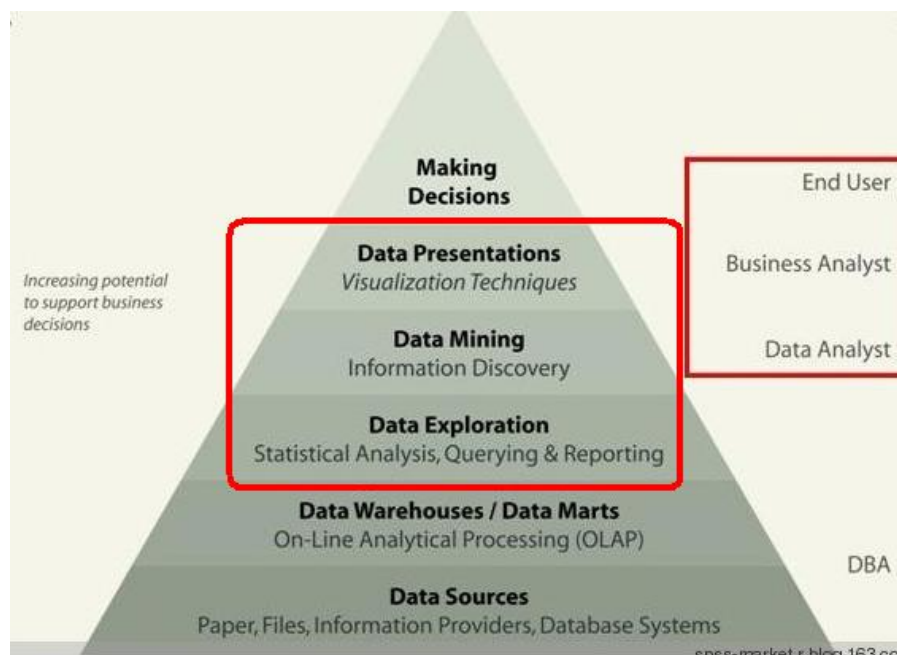
集中趋势指标	均值(mean)	<ul style="list-style-type: none"> 即平均数，$mean = 1/n * \sum(X1:Xn)$； 均值能够利用所有已知信息，但是对异常值(极小或极大值)很敏感；
	中位数(median)	<ul style="list-style-type: none"> 排序后居于中间位置的数值，有序尺度常用； 不能充分利用已知的所有变量信息，但不受异常值的影响；
	众数(mode)	<ul style="list-style-type: none"> 出现最频繁的数值，代表分布中的高峰； 名义尺度(分组数据)常用
变异性指标	极差(range)	<ul style="list-style-type: none"> 最大值与最小值之差，$range = max - min$； 直接受到异常值影响；
	方差(variance)	<ul style="list-style-type: none"> 离均差(观测值与均值之间的差)平方的均值； $var = 1/(n-1) * \sum((Xi - mean)^2)$； 数据分布越分散(远离均值)，方差越大；
	标准差 (standard deviation)	<ul style="list-style-type: none"> 方差的平方根，$stdev = \sqrt{var}$； 与数据本身有相同的量纲，常用；
变异性指标	偏度(skewness)	<ul style="list-style-type: none"> 刻画数据在均值两侧偏差趋势的差异性 对称分布：skewness=0，mean=median=mode； 右偏分布：skewness>0，mean>median>mode； 左偏分布：skewness<0，mean<median<mode；
	峰度(kurtosis)	<ul style="list-style-type: none"> 衡量分布曲线相对平滑或突起程度 kurtosis=3，正态分布(Norm distribution)； kurtosis>3，分布曲线比正态分布突起； kurtosis<3，分布曲线比正态分布平缓；

<http://spss-market.r.blog.163.com/>

- 数据挖掘是以查找隐藏在数据中的信息为目标的技术，是应用算法从大型数据库中提取知识的过程，这些算法确定信息项之间的隐性关联，并且向用户显示这些关联
- 数据挖掘思想来源：假设检验，模式识别，人工智能，机器学习
- 常见数据挖掘任务：关联分析，聚类分析，孤立点分析等等
- 例：啤酒与尿布的故事
- 例：《Science》的文章《[科学家摸索出大型数据集内的趋势](#)》

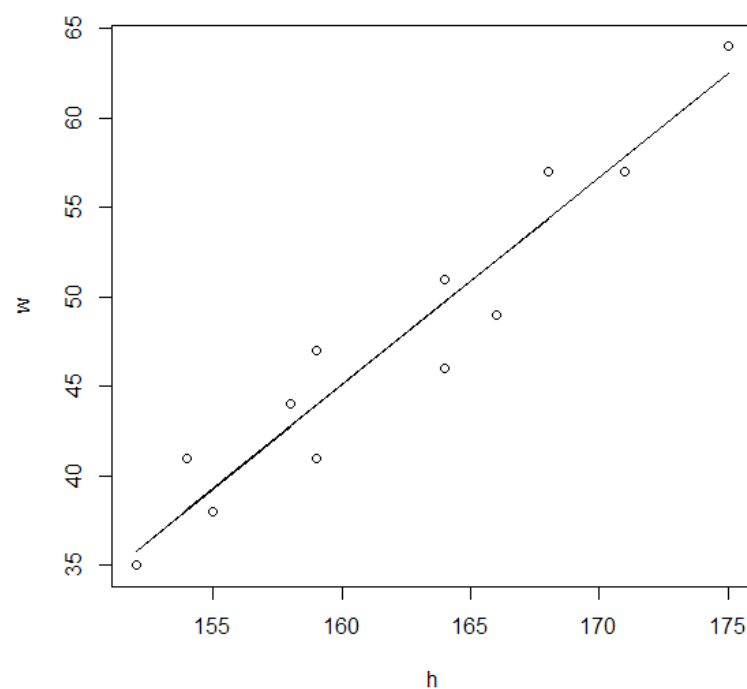
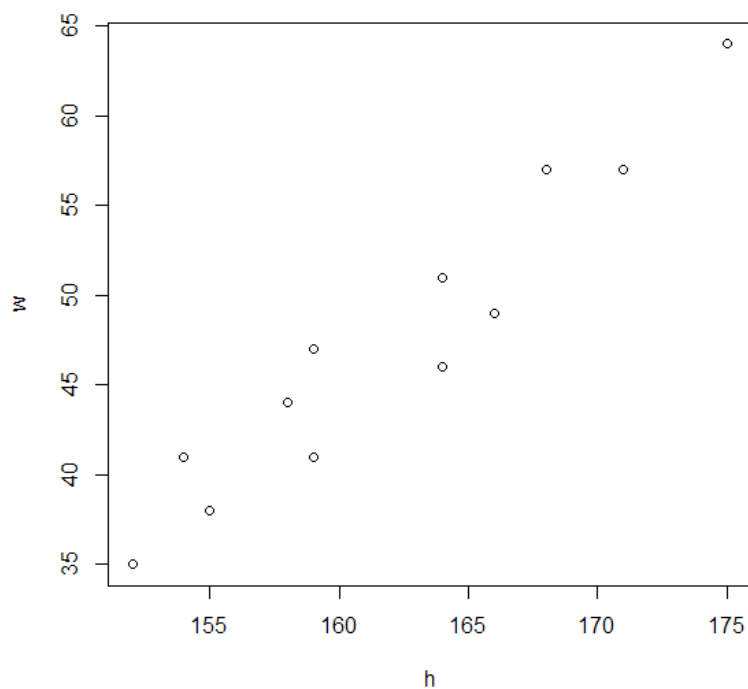


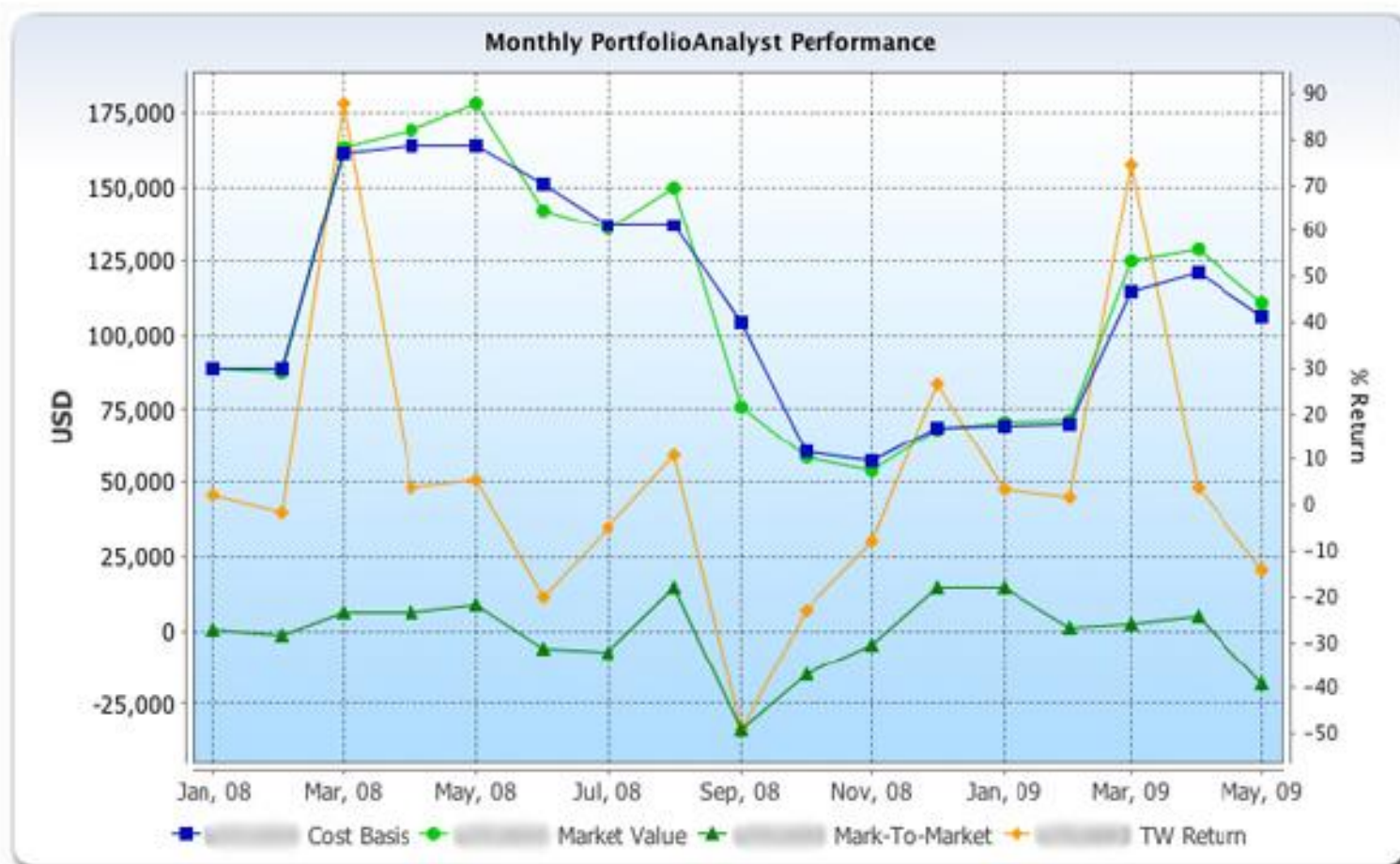
- Business Intelligence , 简称为BI
- BI=数据仓库（存储层）+数据分析和数据挖掘（分析层）+报表（展现层）
- 我们课程的位置

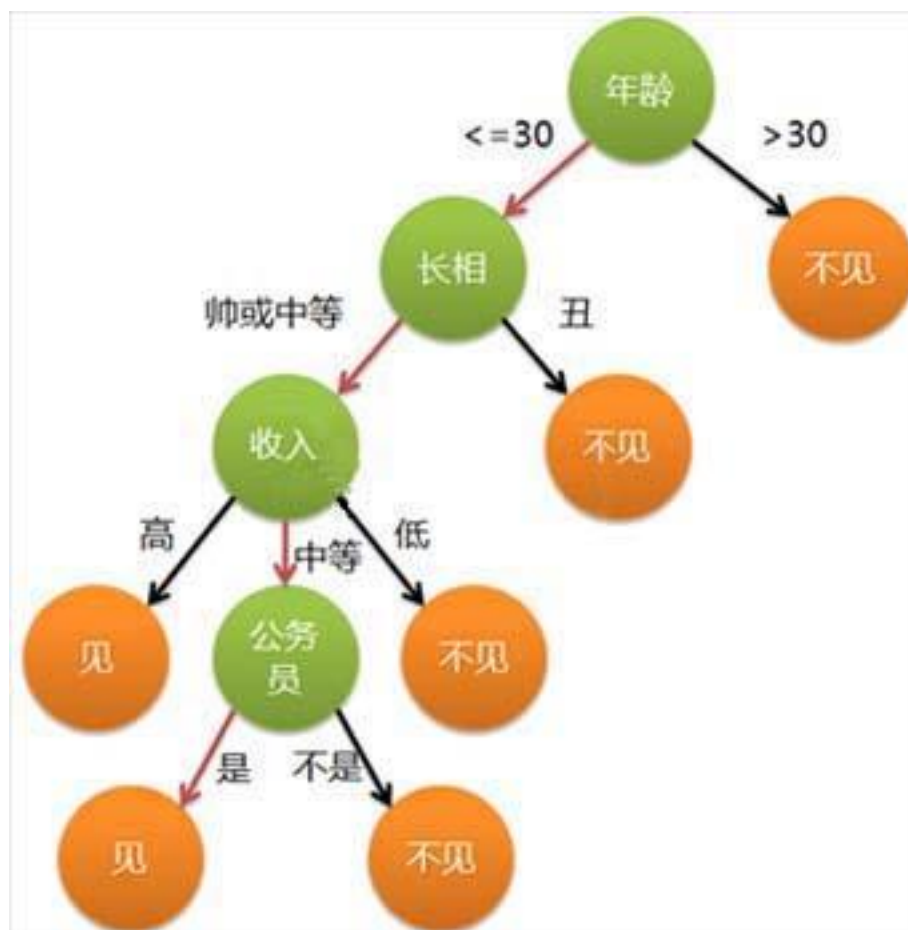


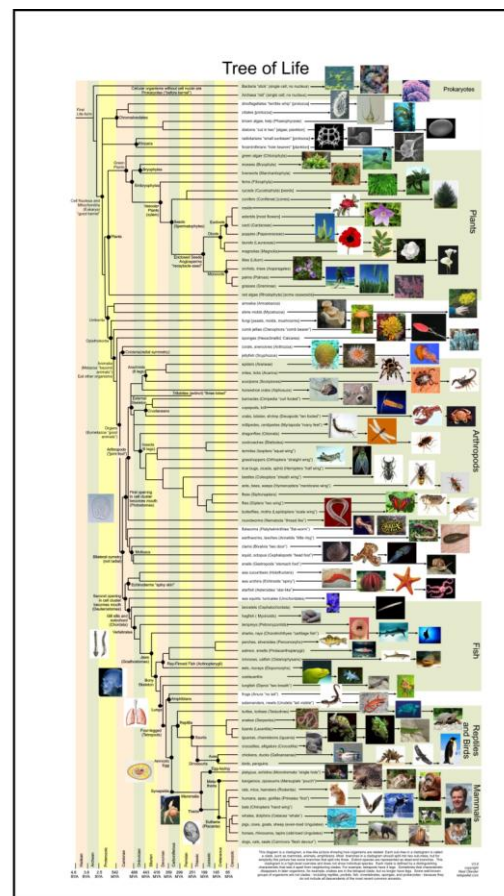
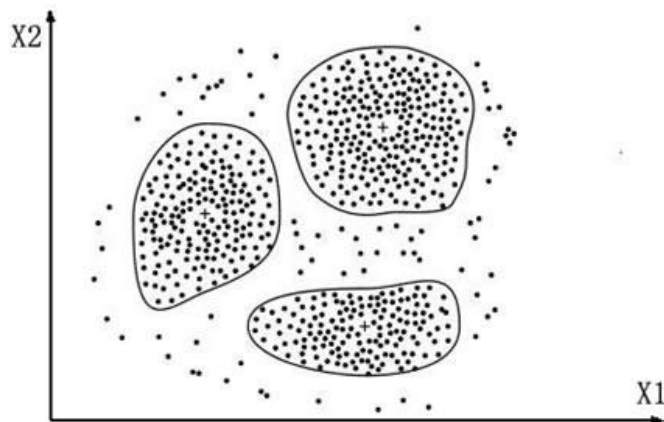
■ 常用算法













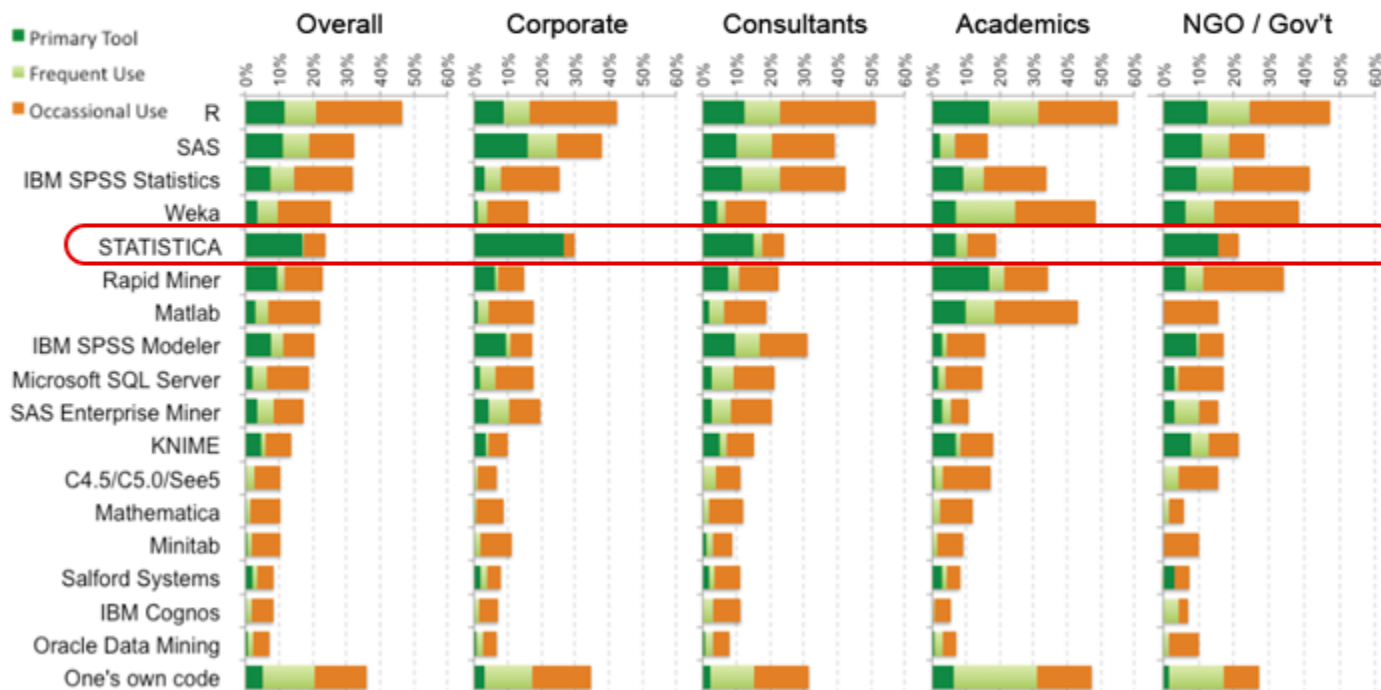
■ 数据分析工具



All Commercial & Open Source Applications

Survey Questions:

- What Data mining/analytic tools did you use in 2010? (rate each as "never", "occasionally", or "frequently")
- What one Data Mining software package do you use most frequently?



- 第一层：业务人员主导
- 第二层：业务人员与数据分析师共同主导
- 第三层：数据分析师主导



- 业务人员制定全部需求及绝大部分分析计划
- 业务人员根据经验感觉提出变量，制定阈值，IT人员查询、汇总数据，数据分析人员制作图表
- 变量的选择，阈值的指定主要根据过往经验，基本属于拍脑袋决策，人脑智能
- 适用于简单，直观的情况，分析技术上相当于最原始简单的决策树（变量少，没有实现最佳分离度，阈值是手工指定的）
- 这是业务人员控制的地盘，数据分析师一般只是起着画图表的作用，业务人员的经验难以替代，数据分析师很容易被替代，因此叫不起价钱
- 这种分析在中国企业里占了大多数

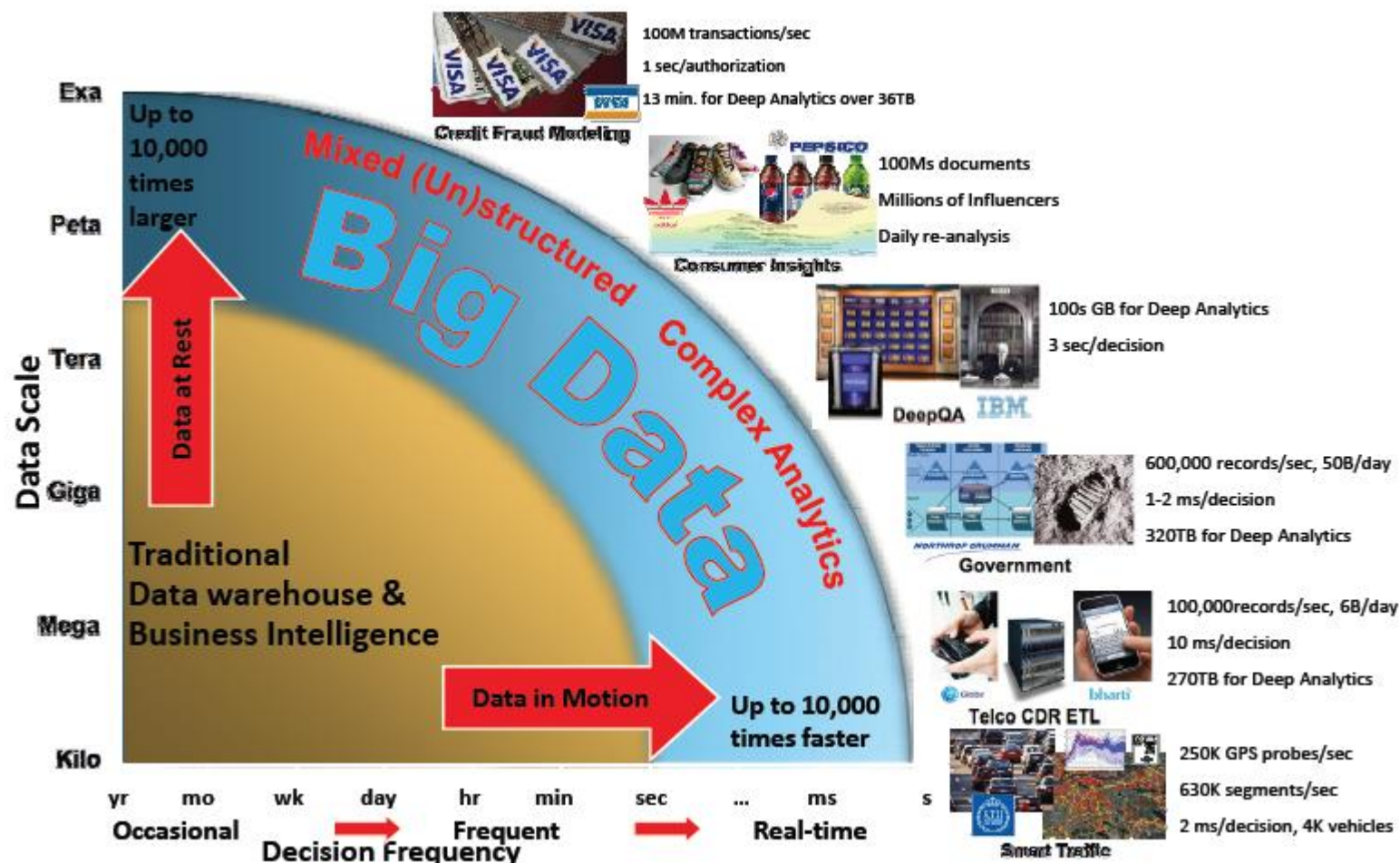
- 复杂度提升。业务人员对单条的样本数据可以用业务经验进行判断。但无法指出全部具体变量，以及变量之间是何种关系。当要对大量样本进行判断时，无法立即利用机器智能
- 数据分析师介入，对数据进行梳理，筛选变量，并建立合适的模型解决问题，最终实现商业智能
- 这一层次，因为业务人员可以肉眼判断，因此肯定是能做的，成功率很高，这也是数据分析师们最喜欢做的层次，能充分体现数据分析师的价值，获得业务人员的肯定

- 业务人员的经验已经用光，完全无法再提供任何有价值的先验知识
- 数据分析师在数据的天空里自由翱翔，运用各种人工智能，模式识别，机器学习的手段，对数据进行挖掘，试图得出有趣的结果
- 这一阶段是数据分析师的理想王国，但失败率非常高
- 由于业务人员基本无法参与，因此在心理上容易产生反感和抗拒

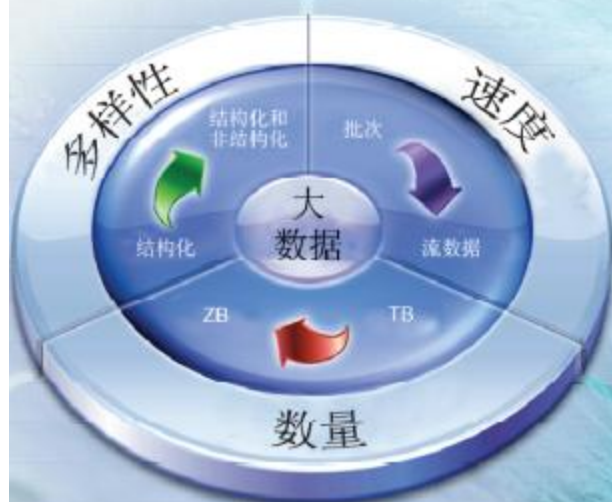


- 第一层（地表的金子）->第二层（浅表的金子）->第三层（深埋的金子）逐层推进
- 如果冒进，容易受到业务人员的抵制和嘲笑
- 业务是数据分析扎根的土壤，提升业务业绩是数据分析的目标，因此任何成功的数据分析都必然和业务紧密相连
- 有意思的分析主题可能比分析技术本身更重要

什么是大数据



从数量庞大、多样化的高速数据中联系上下文提取真知灼见，从而做到以前所不可能做到的事情。



多样性：管理多种关系和非关系数据类型及架构的复杂性

速度：流数据和大批量数据移动

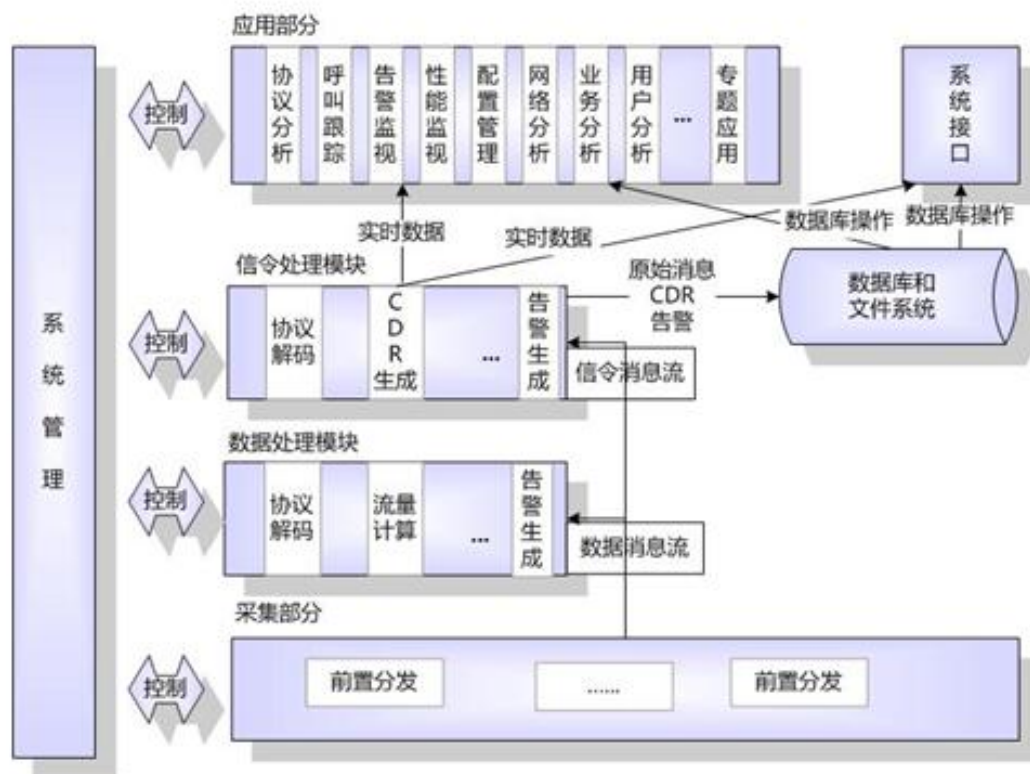
数量：从 TB 扩展到 ZB

- 数据日趋庞大，无论是入库和查询，都出现性能瓶颈
- 用户的应用和分析结果呈整合趋势，对实时性和响应时间要求越来越高
- 使用的模型越来越复杂，计算量指数级上升

传统数据分析工具的困境

- R , SAS , SPSS等典型应用场景为 实验室工具
- 处理数据量受限于内存 , 因此无法处理海量数据
- 使用Oracle数据库等处理海量数据 , 但缺乏有效快速专业的分析功能
- 可以采用抽样等方法 , 但有局限性 , 比如对于聚类 , 推荐系统则无法使用抽样
- **解决方向 : Hadoop集群和Map-Reduce并行计算**

场景一：电信运营商信令分析与监测



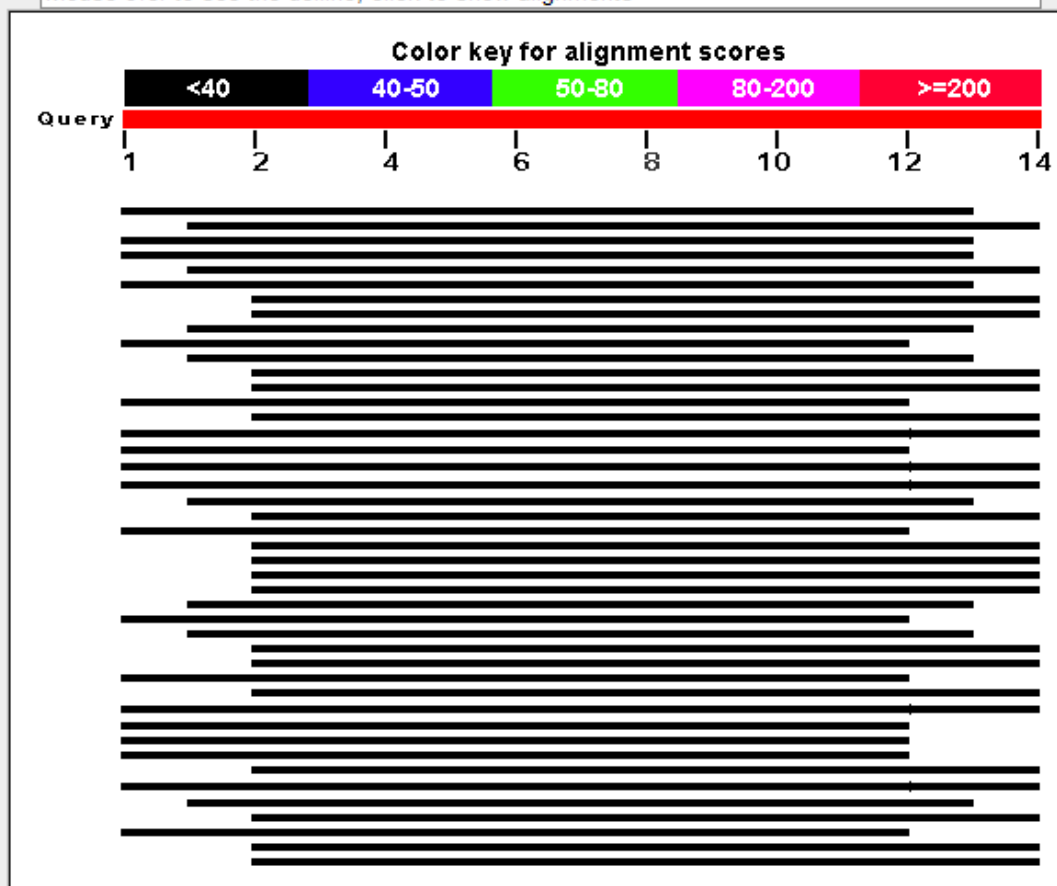
场景一：电信运营商信令分析与监测

- 原数据库服务器配置：HP小型机，128G内存，48颗CPU，2节点RAC，其中一个节点用于入库，另外一个节点用于查询
- 存储：HP虚拟化存储，>1000个盘
- 数据库架构采用Oracle双节点RAC
- 问题：1 **入库瓶颈** 2 **查询瓶颈**

场景二：DNA数据库

Distribution of 50 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



场景三：社会学分析——行为指纹识别

公共指纹

公共语音交往圈	IMSI 1	IMSI 2	IMSI n	总通话时长
A用户IMSI	20%	12%	5%	365
B用户IMSI	15%	13%	2%	310

(0.2, 0.12, ..., 0.05)
(0.15, 0.13, ..., 0.02)

公共短信交往圈	IMSI 1	IMSI 2	IMSI n	月短信条数
A用户IMSI	50%	10%	5%	200
B用户IMSI	20%	13%	2%	260

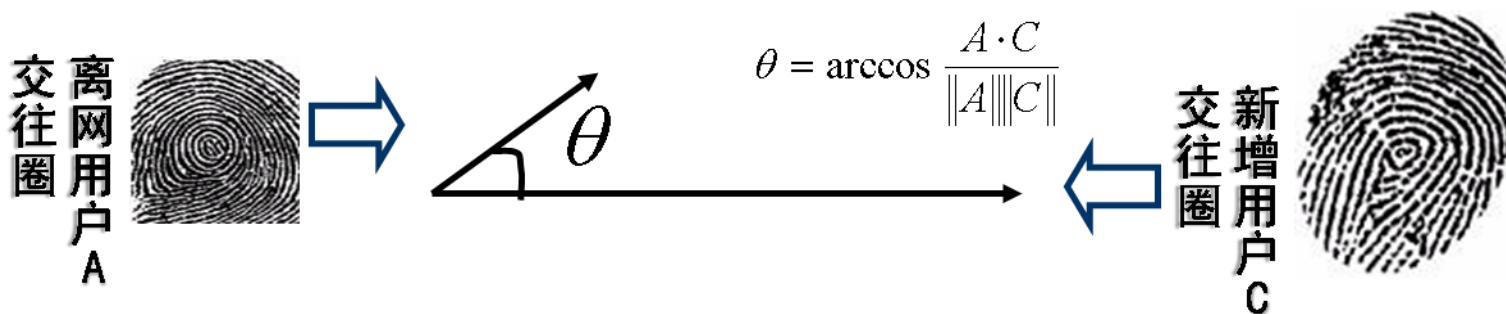
特征向量

(0.5, 0.1, ..., 0.05)
(0.2, 0.13, ..., 0.02)

公共基站	CGI 1	CGI 2	CGI n	关机
A用户IMSI	20%	12%	5%	20%
B用户IMSI	15%	13%	2%	5%

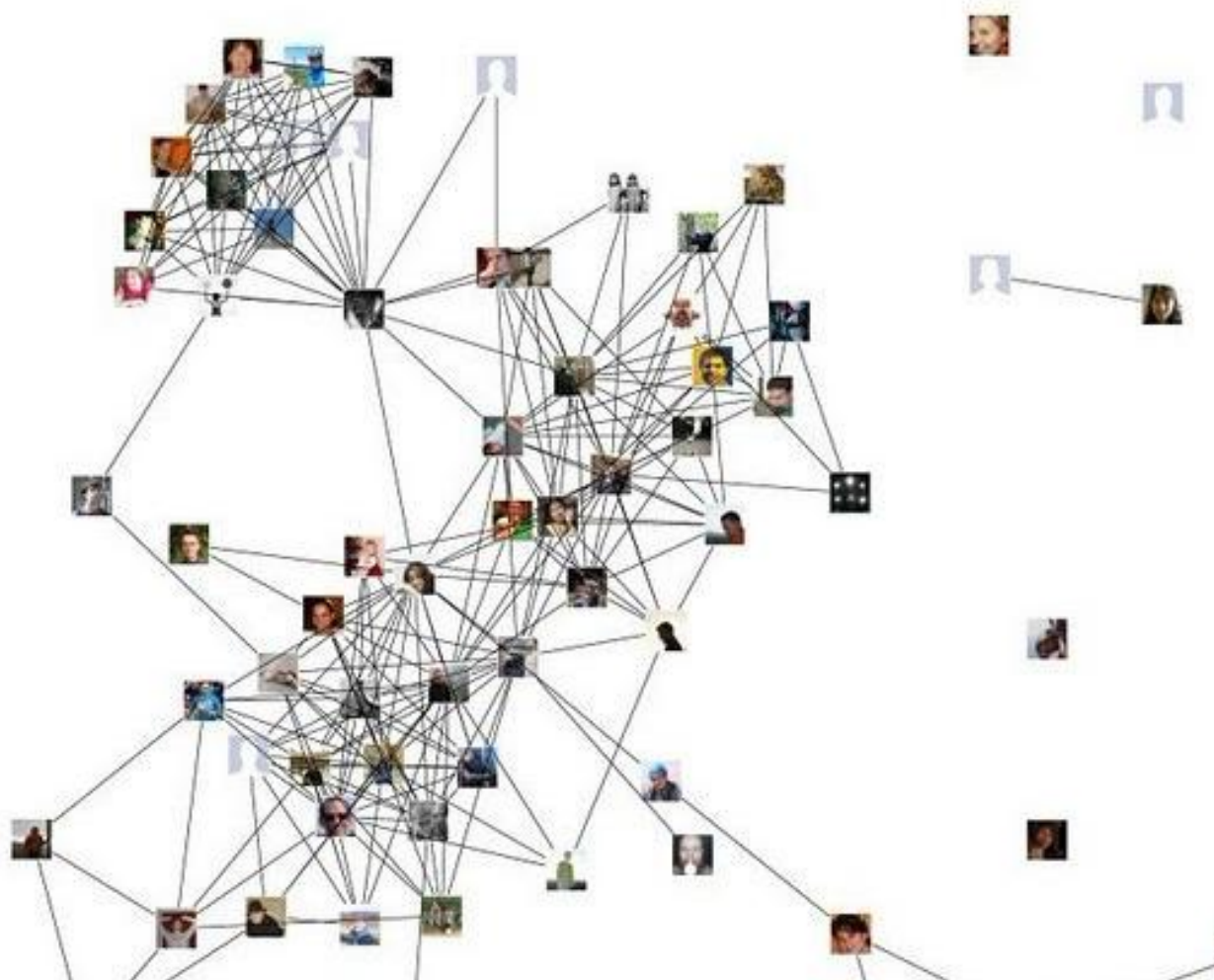
(0.2, 0.12, ..., 0.05, 0.2)
(0.15, 0.13, ..., 0.02, 0.05)

$$\text{sim}(x, y) = \cos(x, y) = \frac{(x, y)}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^n x_i \cdot y_i}{\left(\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i^2 \right)^{1/2}}$$



- 当 θ 为 90° 时，AC两个矢量完全不相关，即两个号码的交往圈相似度最低
- 当 θ 为0 时，AC两个矢量完全相关，即两个号码的交往圈相似度最高
- 当 θ 越接近0，说明两个号码的交往圈越相似

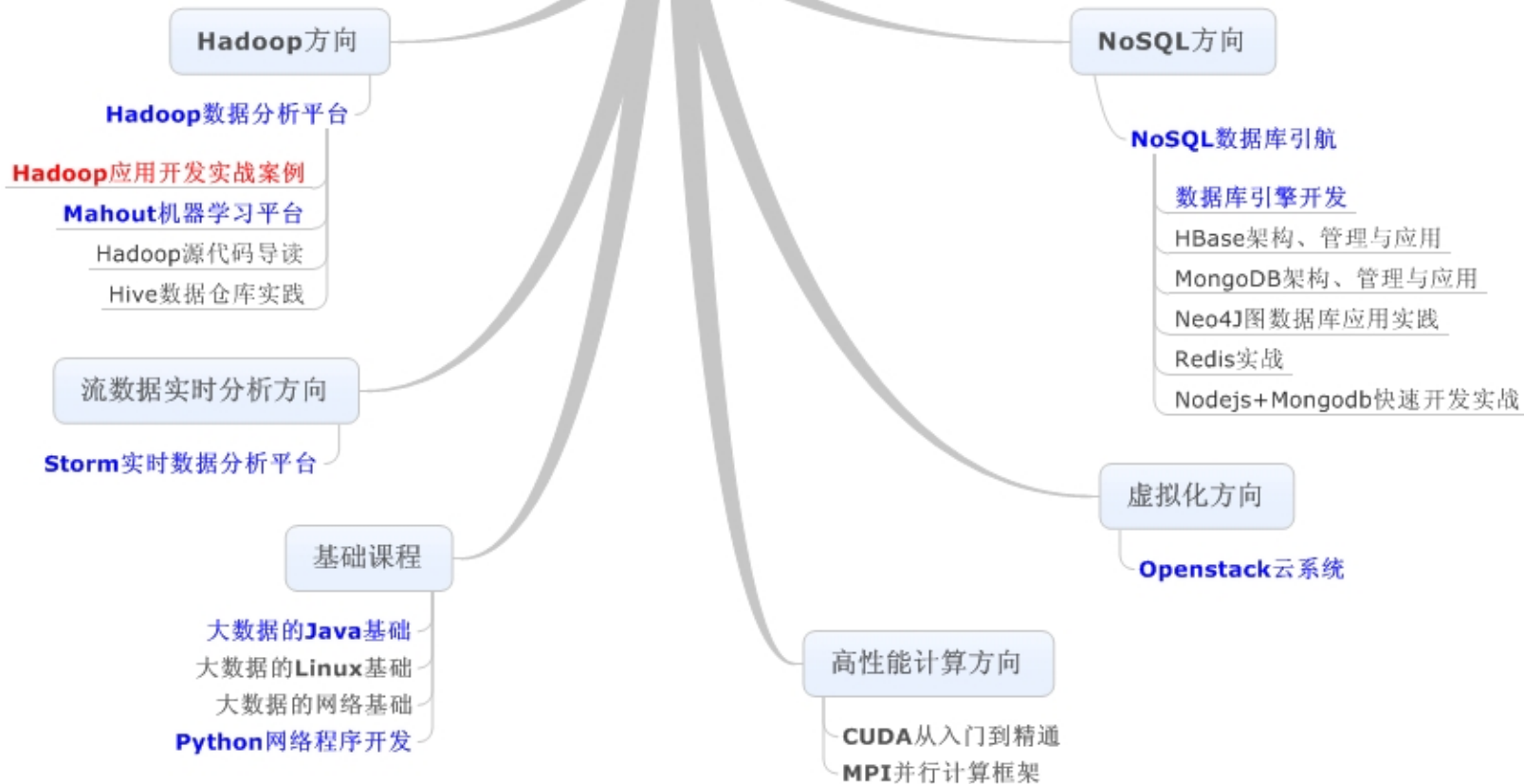
场景三：社会学分析——人物重要度计算

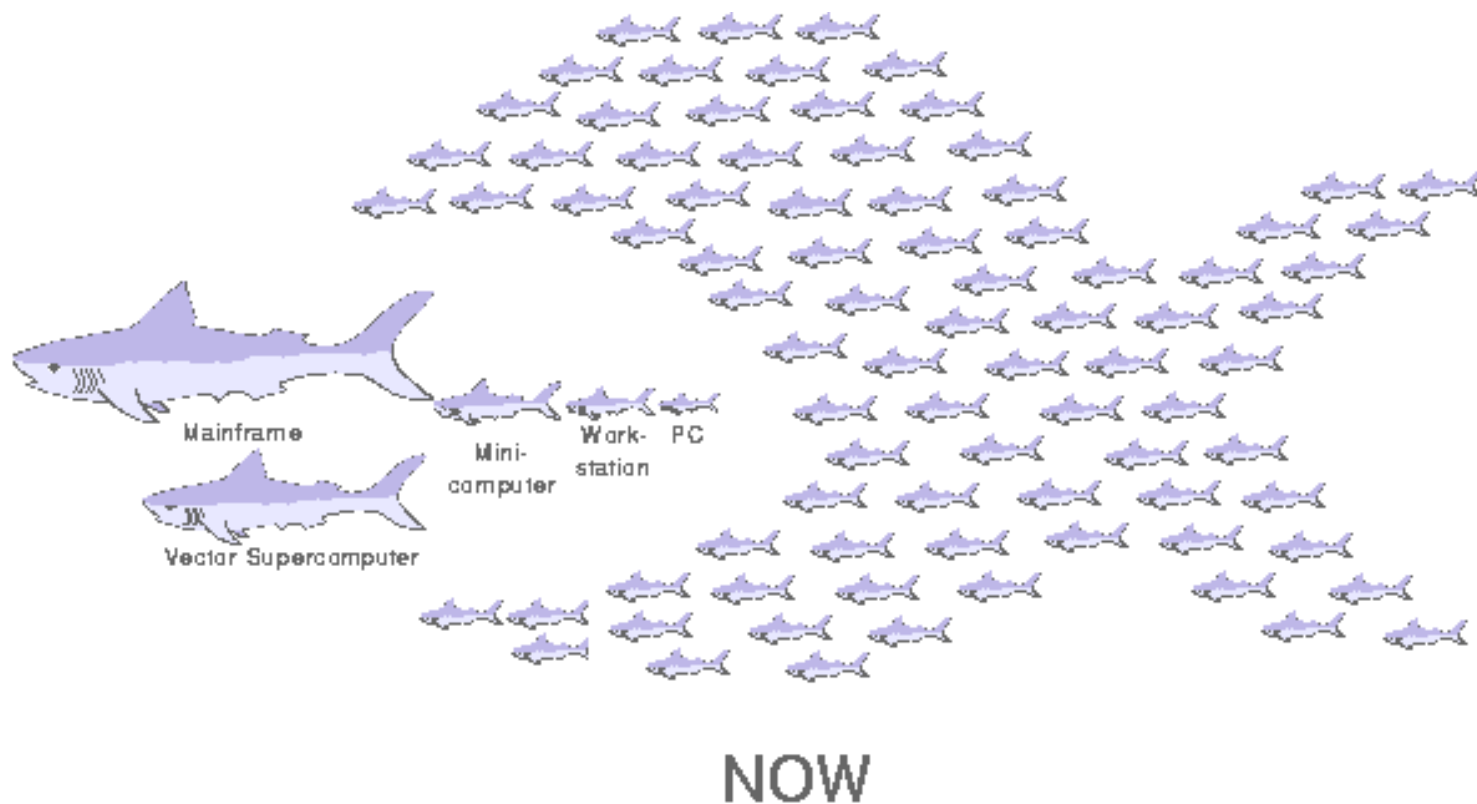


DATAGURU专业数据分析社区

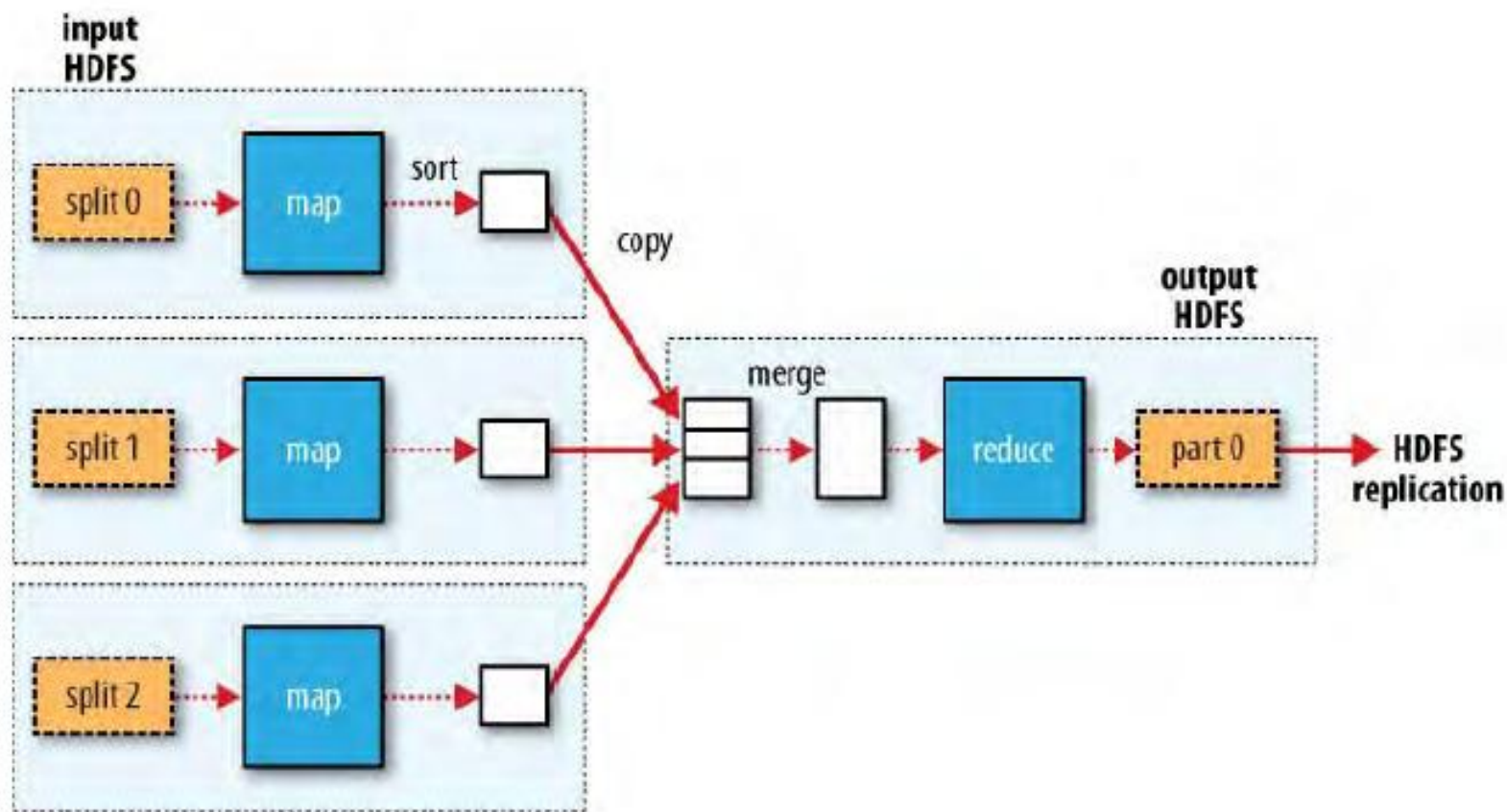
- 完美解决性能瓶颈，在可见未来不容易出现新瓶颈
- 过去所拥有的技能可以平稳过渡。比如SQL、R
- 转移平台的成本有多高？平台软硬件成本，再开发成本，技能再培养成本，维护成本

大数据与云计算方向线路图

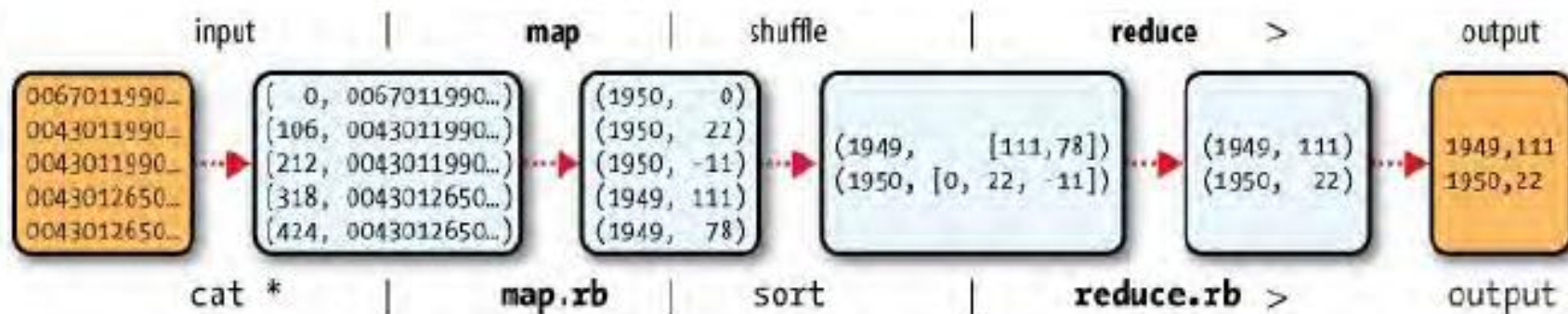




Map-Reduce编程模型



分析气象数据的Map-Reduce程序



	One Iteration	Multiple Iterations	Not good for MapReduce
Clustering	Canopy	KMeans	
Classification	Naïve Bayes, kNN	Gaussian Mixture	SVM
Graphs		PageRank, Connected Components	
Information Retrieval	Inverted Index	Topic modeling (PLSI, LDA)	




Perfect fits



OK fits

small shared info have to be synchronized across iterations (typically through filesystem)



Not good

- require **large shared info** with a lot of synchronization.
- Traditional parallel framework like MPI is better suited for those.

Best for MapReduce:

- ❑ **Single** pass, keys are uniformly distributed.

OK for MapReduce:

- ❑ **Multiple** pass, intermediate states are small

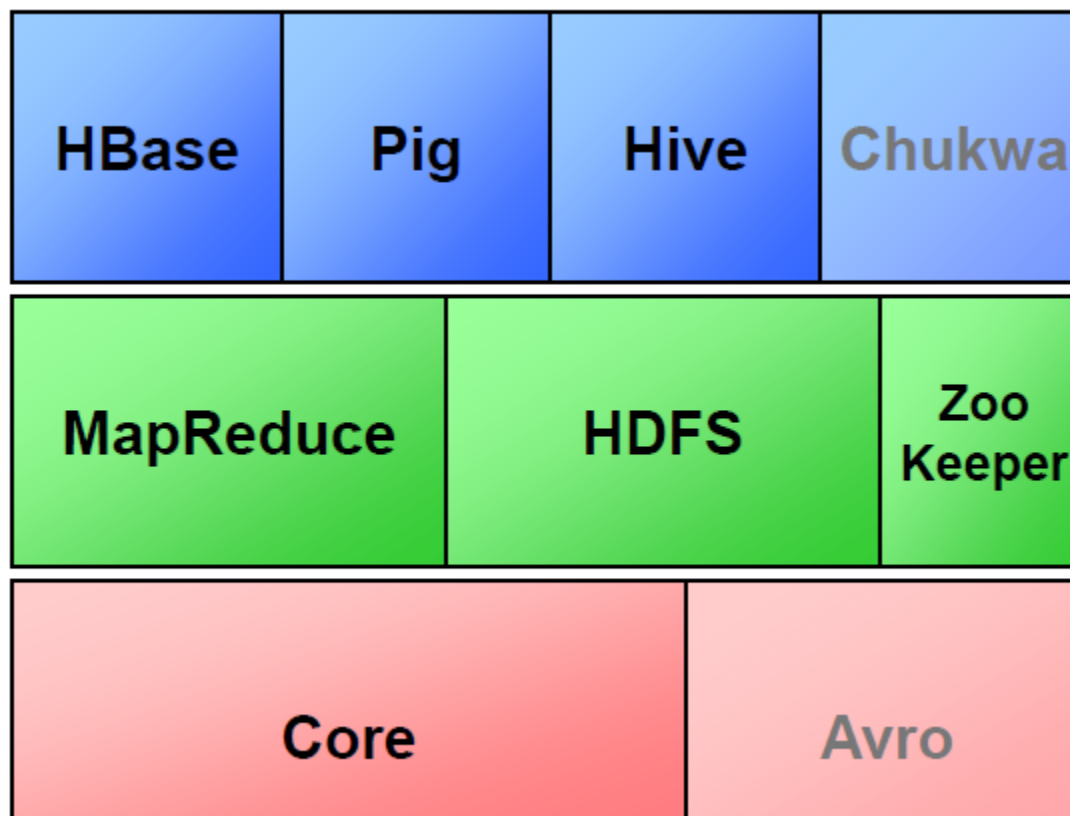
Bad for MapReduce

- ❑ More Data Sharing & More Iterations
- ❑ Fine-grained synchronization is required
- ❑ e.g. SVM, HMM

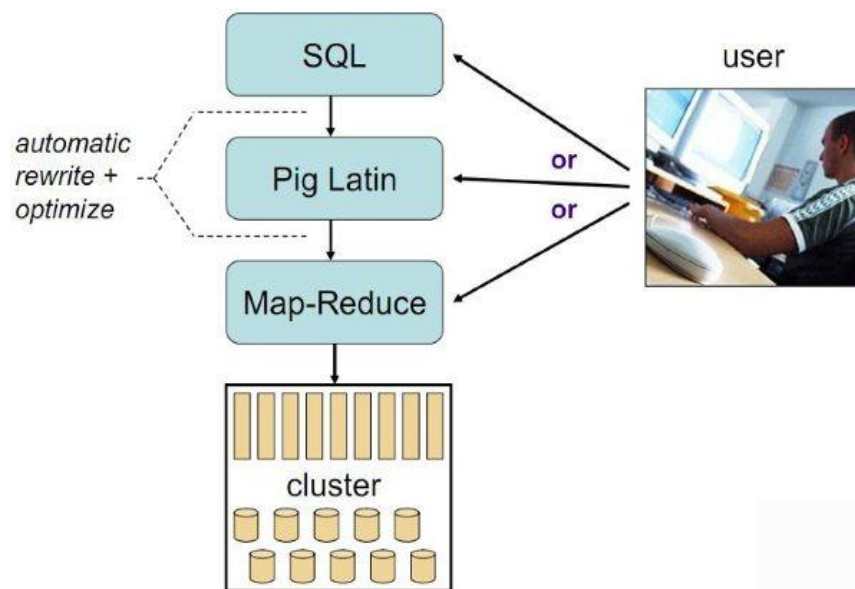
Why not Hadoop ?

- Java ?
- 难以驾驭 ?
- 数据集成困难 ?
- Hadoop vs Oracle

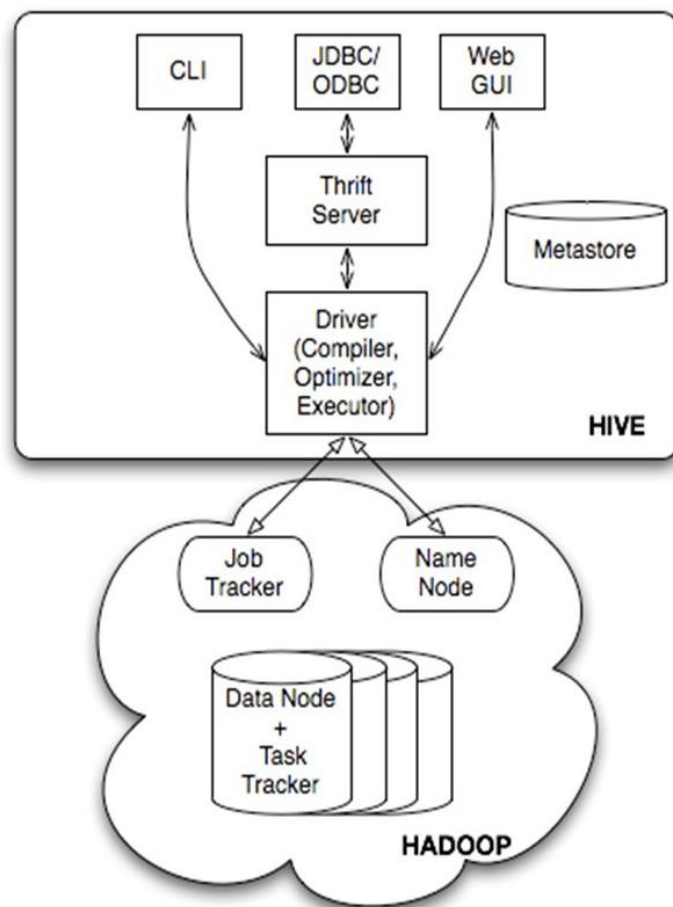
- 主流，Map-Reduce：Java程序
- 轻量级的脚本语言：Pig
- SQL技巧平稳过渡：Hive
- 机器学习平台：Mahout
- NoSQL：HBase



- Pig可以看做hadoop的客户端软件，可以连接到hadoop集群进行数据分析工作
- Pig方便不熟悉java的用户，使用一种较为简便的类似于SQL的面向数据流的语言pig latin进行数据处理
- Pig latin可以进行排序、过滤、求和、分组、关联等常用操作，还可以自定义函数，这是一种面向数据分析处理的轻量级脚本语言
- Pig可以看做是pig latin到map-reduce的映射器



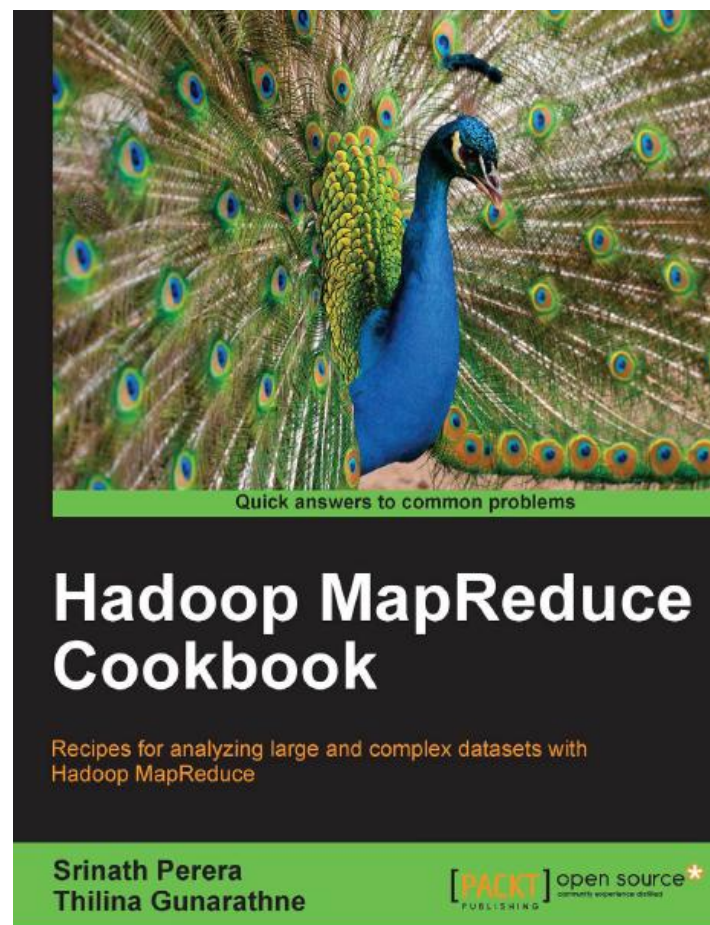
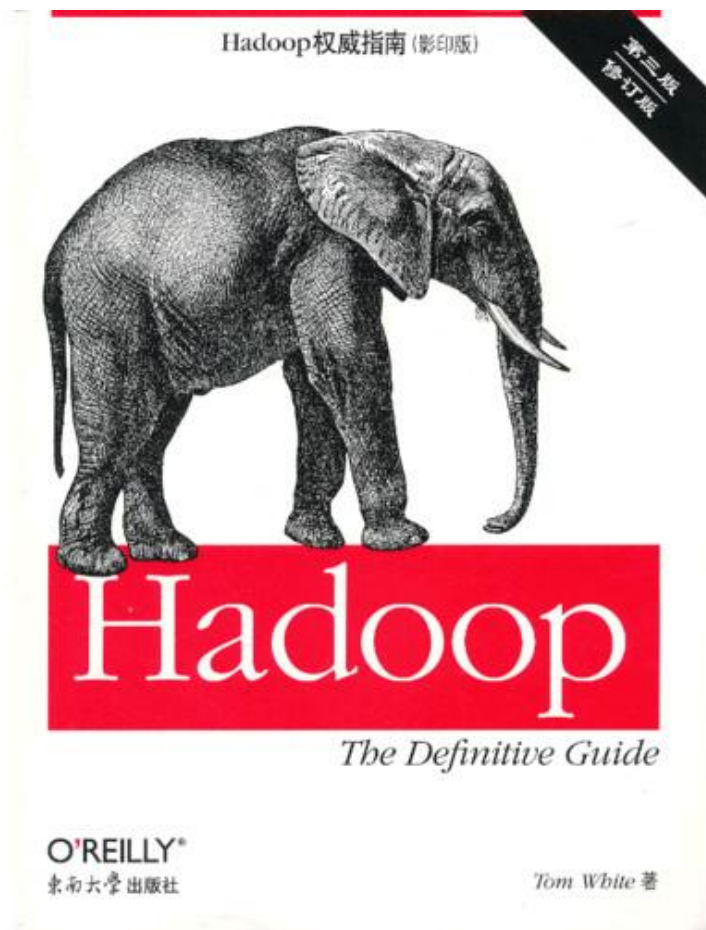
- 数据仓库工具。可以把Hadoop下的原始结构化数据变成Hive中的表
- 支持一种与SQL几乎完全相同的语言HiveQL。除了不支持更新、索引和事务，几乎SQL的其它特征都能支持
- 可以看成是从SQL到Map-Reduce的映射器
- 提供shell、JDBC/ODBC、Thrift、Web等接口



- Mahout的主要目的是实现可伸缩的机器学习算法（就是算法的M-R化），但也不一定要求基于Hadoop平台，核心库中某些非分布式的算法也具有很好的性能
- 目标是帮助开发人员快速建立具有机器智能的应用程序，目前比较成熟和活跃的主要包括
 - 1 频繁模式挖掘
 - 2 聚类算法
 - 3 分类器
 - 4 推荐系统
 - 5 频繁子项挖掘

Mahout currently has

- Collaborative Filtering
- User and Item based recommenders
- K-Means, Fuzzy K-Means clustering
- Mean Shift clustering
- Dirichlet process clustering
- Latent Dirichlet Allocation
- Singular value decomposition
- Parallel Frequent Pattern mining
- Complementary Naive Bayes classifier
- Random forest decision tree based classifier
- High performance java collections (previously colt collections)
- A vibrant community
- and many more cool stuff to come by this summer thanks to Google summer of code



Dataflow Scripting with Hadoop

Programming

Pig



O'REILLY®

Alan Gates

Data Warehouse and Query Language for Hadoop



Programming

Hive



O'REILLY®

*Jason Rutherglen,
Dean Wampler &
Edward Capriolo*

*Building Effective Algorithms and Analytics
for Hadoop and Other Systems*

MapReduce Design Patterns



O'REILLY®

Donald Miner & Adam Shook

← → ↺ 🏠 🌐 www.manning.com/owen/



[Home](#) | [Ordering Info](#) | [Shopping Cart](#)

Keep up with Manning for special offers and updates

✉ Receive our email newsletter

🐦 Follow Manning on Twitter

📘 Become a Manning Facebook fan

Catalog

Java
Microsoft & .NET
All by Title
All by Subject
Mobile Formats

Author Blogs

Jon Skeet
Phillip Trelford
Michael Fogus
Richard Siddaway
Ayende Rahien
more...

Author Calendar

Manning Author Events



Mahout in Action

Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman

October, 2011 | 416 pages
ISBN 9781935182689

ADD TO CART \$44.99 pBook + eBook (includes PDF, ePub, and Kindle)

ADD TO CART \$35.99 eBook Only (includes PDF, ePub, and Kindle)

Browse all our mobile format ebooks.

This eBook includes **audio and video segments**. **Listen and Watch** as the authors share their insights on specific topics.

RESOURCES

Look Inside

- Preface
- About this book
- Table of Contents
- Index

Resources

- [Author Online](#)
- [Multimedia extras](#)
- [Meet Mahout \(Green Paper - PDF\)](#)
- [Clustering Wikipedia articles \(PDF\)](#)
- [Computing Average Log-Likelihood \(PDF\)](#)

Downloads

- [Source code \(102 KB\)](#)
- [Sample chapter 1](#)
- [Sample chapter 8](#)

SUMMARY

Mahout in Action is a hands-on introduction to machine learning with Apache Mahout. Following real-world examples, the book presents practical use cases and then illustrates how Mahout can be applied to solve them. Includes a free audio- and video-enhanced ebook.

ABOUT THE TECHNOLOGY

“A h
of n

典型实验环境（拥有服务器）

- 服务器：ESXi，可以在上面部署多台虚拟机，能同时启动3台
- PC：要求linux环境或windows+Cygwin，linux可以是standalone或者使用虚拟机
- SSH：windows下可以使用SecureCRT或putty等ssh client程序，作用是用来远程连接linux服务器，linux下可以直接使用ssh命令
- Vmware client：用于管理ESXi
- Hadoop：使用1.x或2.x

典型实验环境（只有PC或笔记本，基于win）

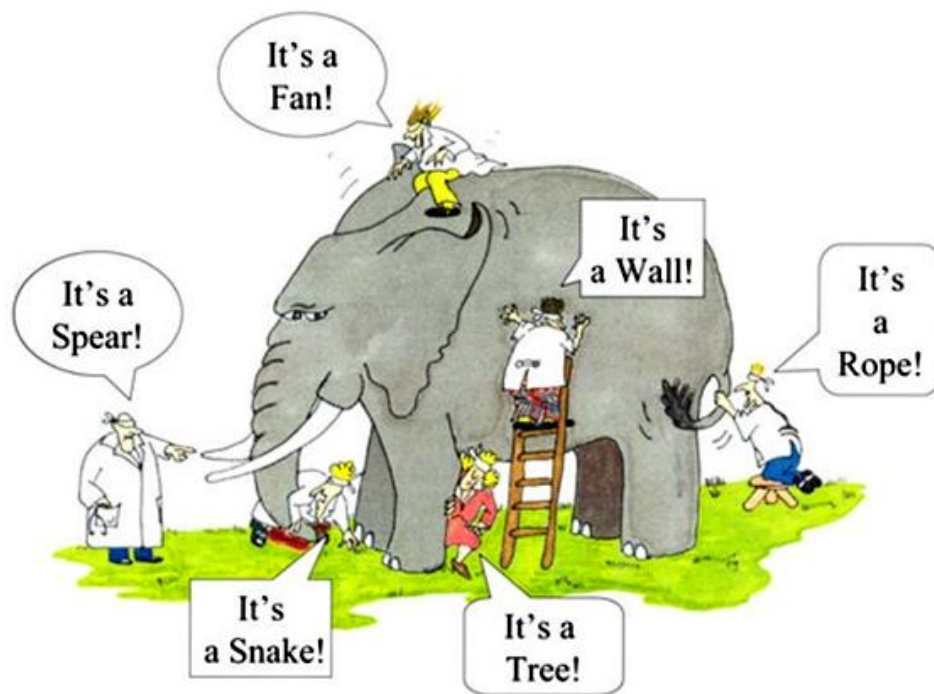
- 至少4G内存，最好运行64位windows系统，因为32位xp只能支持3G多的内存
- 安装vmware workstation或virtual box
- 部署3台虚拟机，能同时运行，如果只能运行2台虚拟机，那么可以把host也作为一个节点（使用cygwin），虚拟网络配置为网桥方式
- 安装linux和java
- 如果配置实在太低只好使用伪分布式

- 部署Pig
- 部署Hive
- 部署Mahout

课程内容：案例列表（初定）

- 巨型网站日志系统分析，提取KPI数据(Map-Reduce)
- 电信运营商LBS应用，分析手机用户移动轨迹(Map-Reduce)
- 电信运营商用户分析，通过通话指纹判断重入网用户(map-Reduce)
- 电子商务推荐系统设计(Map-Reduce)
- 更复杂的推荐系统场景(Mahout)
- 社交网络，判断微博用户关系亲疏程度，发现社区(Pig)
- 在社交网络中衡量节点的重要程度(Map-Reduce)
- 聚类算法应用，分析优质客户(Map-Reduce,Mahout)
- 金融数据分析，从历史数据中提取逆回购信息(Hive)
- 通过数据分析制定股票策略(Map-Reduce,Hive)
- GPS应用，签到数据分析(Pig)
- Map-Reduce全排序实现和优化
- 中间件开发，让多个Hadoop集群协作起来

- 参与授课的几位一线工程师：张丹，黄俊，郑梓力
- 全视角：甲方业务人员需求（**业务目标**），数据分析师（**算法设计**），数据架构师/ETL工程师(ETL过程设计与控制)，IT架构师/运维（IT基础平台设计与部署），程序员（**代码实现**），信息展现工程师（数据展现）
- 会作一定的简化（比如略去非技术因素，问题数据整理等更耗时的工作）



- **Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。**
- **关于逆向收费式网络的详情，请看我们的培训网站 <http://edu.dataguru.cn>**



Thanks

FAQ时间