浅谈大数据及其相关技术

马 帅

北京航空航天大學
BEIHANG UNIVERSITY

个人简介

- 北京航空航天大学计算机学院教授、博士生导师；中国计算机学会数据库专业委员会委员，大数据专家委员会委员。
- 2011年作为海外优秀中青年人才加入北京航空航天大学计算机学院软件开发环境国家重点实验，并特聘为教授。
- 获得了北京大学(2004)和英国爱丁堡大学 (2011)的两个博士学位。英国爱丁堡大学博士后，并曾在美国贝尔实验室总部实习，在微软亚洲研究院访问。

Homepage: http://mashuai.buaa.edu.cn

Email: mashuai@buaa.edu.cn

Address: Room G1122,
New Main Building,
Beihang University

北航大数据科学与工程国际联合研究中心

- International Research Centre on Big Data (RCBD)
  - Founded in September, 2012.
  - Led by **Prof. Wenfei Fan** (ACM Fellow, Fellow of the Royal Society of Edinburgh, Scotland) .
- Research Topics
  - Big Data Analysis: Theory and Applications
  - Data Quality: The Other Side of Big Data
  - Querying Big Data beyond MapReduce
  - Querying Big Social Data

973 Grant on Big Data at Beihang

- 网络信息空间大数据计算的基础研究(2014-2018)
  - Chief Scientist: Prof. Jinpeng Huai.
  - 8 institutes involved
  - Focus on “computing theory and practice on Big Data”
  - http://cnbigdata.org/

北京航空航天大學 BEIHANG UNIVERSITY

國防科学技術大学 National University of Defense Technology

ISCAS

HUAWEI

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CNCERT/CC 国家计算机网络应急技术处理协调中心

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Baidu 百度

提纲

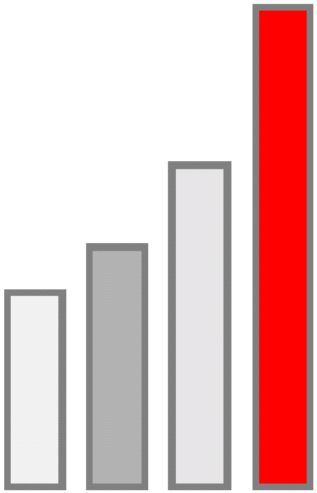
- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据相关技术

大数据(维基百科)

- [英文定义] **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- [中文定义]**大数据**或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息[1][2]。
- 在总数据量相同的情况下，与个别分析独立的小型数据集相比，将各个小型数据集合并后进行分析可得出许多额外的信息和数据关系性，可用来察觉商业趋势、判定研究质量、避免疾病扩散、打击犯罪或测定实时交通路况等；这样的用途正是大型数据集盛行的原因[3][4][5]

- [1]Kusnetzky, Dan. What is "Big Data?". ZDNet.
- [2]Vance, Ashley. Start-Up Goes After Big Data With Hadoop Helper. New York Times Blog. 2010.
- [3]Data, data everywhere. The Economist. [2010-02-25].
- [4] Cat Casey and Alejandra Perez. E-Discovery Special Report: The Rising Tide of Nonlinear Review. Hudson Global. [1 July 2012]
- [5]What Technology-Assisted Electronic Discovery Teaches Us About The Role Of Humans In Technology — Re-Humanizing Technology-Assisted Review. Forbes. [1 July 2012]

“大数据”特征 – 4V

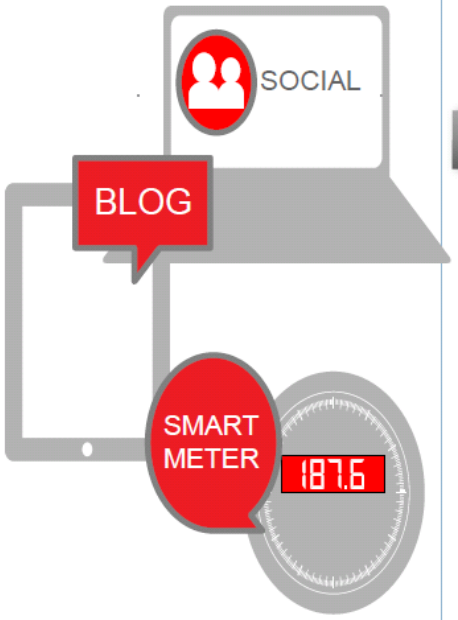


VOLUME

规模大

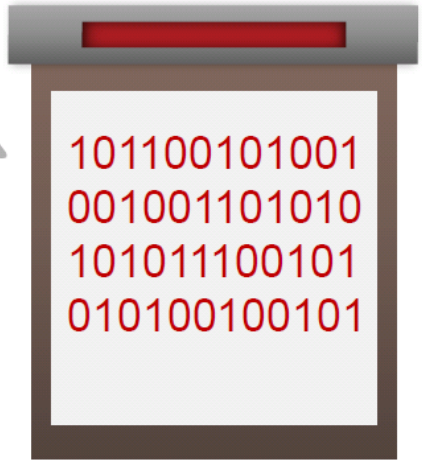
VELOCITY

变化快



VARIETY

种类杂



VALUE

价值密度低

提纲

- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据带来的挑战

“大数据”溯源

- 2008年9月4日《Nature》刊登了一个名为“Big Data”的专辑
  - Researchers need to adapt their institutions and practices in response to torrents of new data — and need to complement smart science with smart searching.
    http://www.nature.com/news/specials/bigdata/
- 2009年10月微软为纪念Jim Gray, 出版了“第四范式—数据密集的科学发现 (The Fourth Paradigm — Data Intensive Scientific Discovery)
- 2011年2月11日：Science刊登了名为Dealing with Data的专辑，联合Science: Signaling、Science: Translational Medicine和Science Careers推出相关专题，讨论数据对科学研究的重要性

“大数据”溯源

- 2012年3月29日，美国总统科技政策办公室OSTP（Office of Science and Technology Policy）宣布了每年投资两亿美元的“大数据研究计划”（Big Data R&D Initiative）
- 同天，我国科技部发布的“‘十二五’国家科技计划信息技术领域2013年度备选项目征集指南”把“大数据研究”列在首位

“大数据”溯源

- 美国国防部高级研究计划局(DARPA)项目:
  - 网络内部威胁计划通过分析图像和非图像的传感器信息和其他 来源的信息，进行网络威胁的自动识别和非常规的战争行为。
  - 多尺度异常检测项目解决大规模数据集的异常检测和特征化。
  - Machine Reading 项目旨在实现人工智能的应用和发展学习系 统，对自然文本进行知识插入。
  - Mind‘s Eye 项目旨在建立一个更完整的视觉智能。
- 美国国家人文基金会(NEH) 项目：
  - 分析大数据的变化对人文社会科学的影响，如数字化的书籍和 报纸数据库，从网络搜索，传感器和手机记录交易数据。

“大数据”溯源

- 美国能源部 (DOE) 项目：
  - 生物和环境研究计划，大气辐射测量气候研究设施
  - 系统生物学知识库对微生物，植物和环境条件下的生物群落功 能的数据驱动的预测。
- 美国国家科学基金会(NSF) 项目：
  - 基础理论与技术：
    - ✓ 推进大数据科学与工程的核心技术，旨在促进从大量、多样、 分散、异构的数据集中提取有用信息的核心技术。
    - ✓ 深入整合算法，机器和人，以解决大数据的研究挑战。
    - ✓ 开发一种以统一的理论框架为原则的统计方法，可伸缩的网络 模型算法，以区别适合随机性网络的方法
    - ✓ 形成一个独特的学科包括数学、统计基础和计算机算法
  - 开放科学网格(OSG)
    - ✓ 使得全世界超过 8000 名的科学家合作进 行发现，包括寻找希格斯玻色子（“上帝粒子”，宇宙中所有物 质的质量之源）

“大数据”溯源

- 2014年我国科技部关于发布国家重点基础研究发展计划：大数据计算的基础研究
- 面向网络信息空间大数据挖掘的需求，结合1-2种重要应用，研究多源异构大数据的表示、度量和语义理解方法，研究建模理论和计算模型，提出能效优化的分布存储和处理的硬件及软件系统架构，分析大数据的复杂性、可计算性与处理效率的关系，为建立大数据的科学体系提供理论依据。

“大数据”溯源

- 2015年我国科技部国家重点基础研究发展计划： 城市大数据的计算理论和方法
- 面向公共安全领域以及智能城市的实际需求，研究空间信息数据、社会网络数据等的协同表示，研究面向信息空间、物理世界和人类社会三元空间的协同感知与群智认知理论，提出视觉计算模型，建立深度计算模型，研究三元空间虚拟交互与智能控制新的模式，适应社会管理、智能城市和工业化生产等方面应用需求。

“大数据”溯源

- **国家自然科学基金委(2014)：**
  - **大数据技术和应用中的挑战性科学问题研究**
- 海量、异构和混杂大数据的广泛存在与爆炸式增长给当代信息传输、存储、计算以及面向各种应用的数据处理技术提出了前所未有的挑战。如何根据社会与国家发展需求，高效准确地传输、存储与计算各种大数据、并从已存在或动态变化的大数据中挖掘有价值的知识成为亟待解决的科学问题。
- 本重点项目群要求各申请团队：结合具体应用，突破传统研究方法的思维定式，研究和发展革命性的、可满足时代需求的大数据传输、存储、计算和处理的新方法和新技术；主要研究成果需在特定大数据集上得到验证。本重点项目群涉及如下研究方向：
- 面向大数据的知识表达、推理及在线学习理论与方法（F02）
- 基于认知计算的大数据分析方法（F02）
- 面向大数据的粒计算理论与方法（F02）
- 大数据环境下复杂多媒体内容分析、推送与展示（F02）
- 大数据管理系统评测基准的理论与方法（F02）
- 多层多域网络化大数据的高效传输理论与方法（F03）
- 大数据高效能存储与管理方法（F03）
- 大数据高时效计算体系结构与关键技术（F03）
- 大数据结构与关系的发现与简约计算方法（F03）
- 基于大数据的复杂系统行为预测与控制（F03）

“大数据”溯源

- **国家自然科学基金委(2015)：**
- **信息科学部(重点群3个+重点项目若干)**
  - G1. 分布式水声网络定位与探测基础研究（F010701）
  - G2. 网络空间智慧搜索基础研究（F020511）
  - G3. 流程工业知识自动化系统设计方法与应用验证（F030102）

  - 20. 大规模并发系统的理论模型与模型检验（F020101）
  - 24. 面向大数据内存计算的新型计算机体系结构（F020302）
  - 22. 大规模复杂关联数据管理的理论与方法（F020204）
  - 47. 大规模知识关联和文本语义计算方法及应用验证（F0304）
  - 32. 混杂数据的模式识别及敏感内容挖掘理论与方法（F020508）
  - 21. 数据中心资源利用率敏感的编程与编译技术（F020202）
  - 33. 面向图像序列的深度学习理论与方法（F020509）
  - 38. 基于数据与机理分析的有源配电网状态估计与趋优协调控制（F0301）
  - 50. 基于神经可塑性的人和智能假肢融合基础理论与关键技术（F0306
  - 52. 基于多感觉脑认知机制的多模态计算模型与实验验证（F0307）
  - 26. 基于多源数据的可视模型与环境构建及其动态仿真
  - 30. 大规模在线教育群体协同学习与个性化智能导学机理与方法（F020507）
  - 31. 大规模在线教育资源汇聚与组织的理论与方法（F020507）
  - 42. 基于大数据和领域知识的复杂石化过程能效评价与系统优化（F0302）
  - 48. 泛在信息制造环境下的机器人群智计算模型与优化方法（F0305）

“大数据”溯源

- **国家自然科学基金委(2015)：**
- **管理科学部(重点群2个+重点项目若干)**
  - 重点群
    - ✓ “物联网环境下的管理理论与方法”重点项目群（G0103，G0109，G0112）
    - ✓ “大数据驱动的管理与决策若干基础问题研究”重点项目群（G02）
  - 1. 智能健康信息服务管理（G0109，G0112）
  - 2. 基于顾客心理和行为的服务价值度量（G0108）
  - 3. 大数据环境下的智慧制造组织模式和运营管理（G0110）
  - 4. 社会网络中企业舆情管理的理论与方法（G0112）
  - 5. 个体和群体选择行为的实验研究及复杂性分析（G0104，G0109）
  - 6. 面向社会网络的企业产品促销、定价和库存管理研究（G0103）
  - 7. 电子商务中的定向广告模式和运用策略研究（G0103，G0112）
  - 11. 大数据环境下的金融风险传导与防范研究（G0206）
  - 13. 移动互联网环境下的用户行为与商业模式创新（G0208）
  - 16. 网络环境下创业行为与决策机制研究（G0215）
  - 8. 新型城镇化导向下的综合交通管理问题研究（G0103，G0109）
  - 20. 巨灾型突发事件医学应急救援风险分析及机制研究（G0310）
  - 22. 大数据环境下知识融合理论、方法与实现路径研究（G0314）

“大数据”溯源

- **国家自然科学基金委(2015)：**
- **地球科学部(重点项目若干)**
  - **4. 天气、气候与大气环境变化的过程与机制**
    - ✓ （1）重要大气现象中关键变量探测的理论与方法
    - ✓ （2）大气探测资料与其他地球观测资料的集成和应用
    - ✓ （3）天气、气候数值模式的关键过程与技术
    - ✓ （6）区域大气污染机制和数值模拟
  - **5. 全球环境变化与地球圈层相互作用**
    - ✓ （1）亚洲季风系统过去、现在和未来演变的机理
    - ✓ （2）典型暖期亚洲重要气候事件及其机制
  - **11. 对地观测及其信息处理**
    - ✓ （2）泛在地理信息集成与质量评价方法
    - ✓ （3）地理空间大数据表达与管理的理论与方法
    - ✓ （4）地理计算与时空分析的新理论与新方法
    - ✓ （5）地理信息服务新理论与新方法
    - ✓ （7）特殊地物多维波谱库建立及深度挖掘的理论与方法
- **数理科学部(重点项目若干)**
  - 复杂推理的逻辑基础及其量化模型（A0115）
  - 随机树与随机图（A0110）
  - 网络设计中的图论方法（A0116）
  - 复杂多源异构数据协同计算的数学理论与方法（A0117）
  - 复杂系统动力学建模、分析与控制（A0202）
  - 人类健康与医学中的生物力学问题（A0205）
  - 复杂力学问题的计算方法与软件（A02）

“大数据”溯源

- **国家自然科学基金委(2015)：**
- **生命科学部(重点项目若干)**
  - 蛋白质及复合物的结构、修饰和功能调控（C0501）
  - 核酸代谢与基因组稳定性（不包括非编码RNA）（C0502）
  - 感觉信息的神经编码机制（C0904）
- **化学科学部(重点项目若干)**
  - 理论与计算化学中的新方法及应用（B03）
  - 蛋白质检测及其功能研究（B05）
- **医学科学部**
  - 35. 基于生物医学大数据的疾病风险评估与预测方法学研究（H26）
- **工程与材料科学部？**

提纲

- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据带来的挑战

人类 vs. 计算机 + 数据

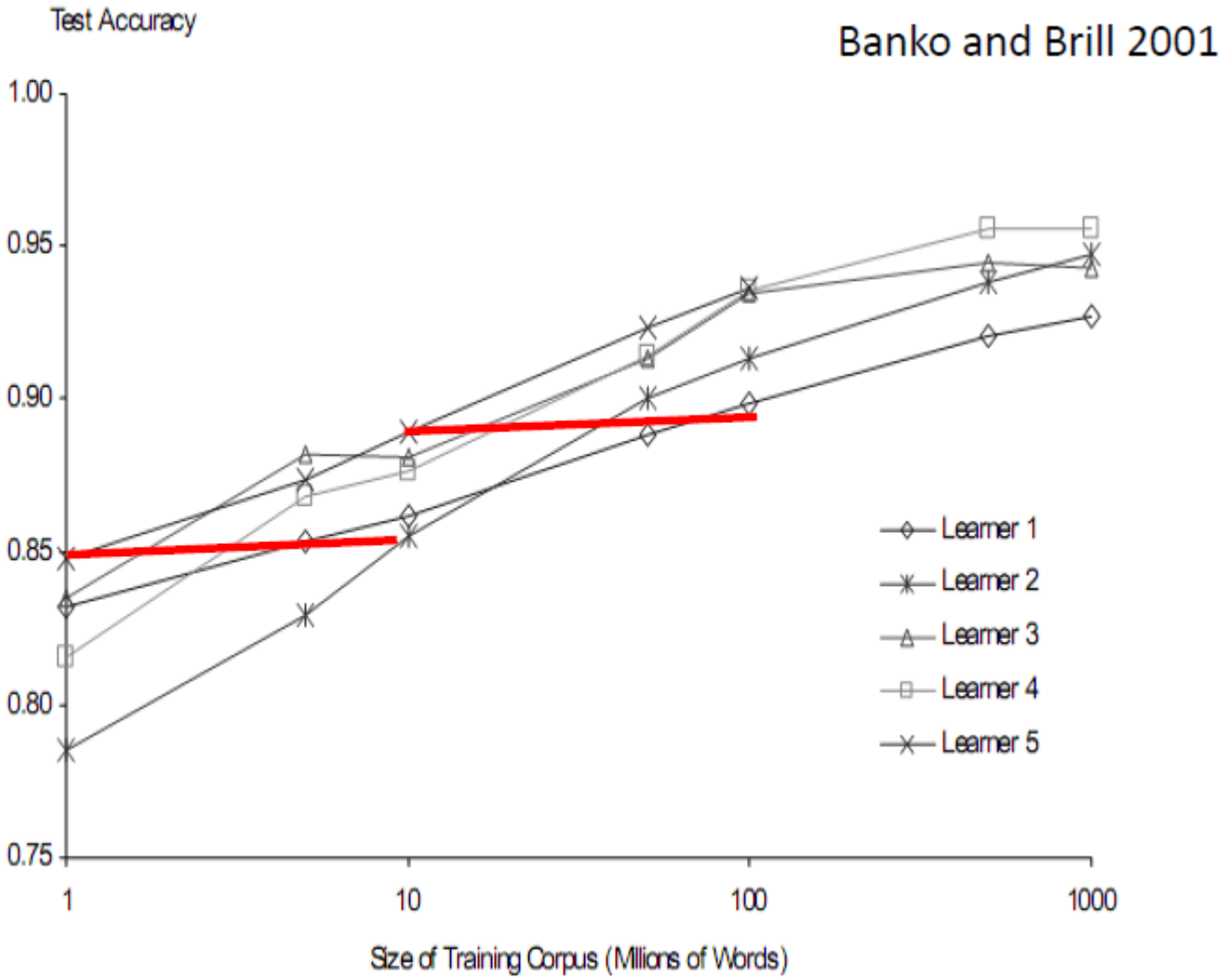
- 2011年2月11日，美国很受欢迎的智力竞答 “危险边缘（Jeopardy）”电视节目
  - IBM的“沃森”系统以绝对优势战胜两名人类顶级选手
  - 和14年前的“深蓝”（战胜**加里·卡斯帕罗夫**）相比，“沃森”除具有超群的计算能力外，更拥有超大规模的数据以及数据处理能力

开普勒第三定律的发现

- 是以太阳为焦点的椭圆轨道运行的所有行星，其椭圆轨道半长轴的立方与周期的平方之比是一个常量。

Planet	Period (yr)	Ave. Dist. (au)	T^2/R^3 (yr^2/au^3)
Mercury	0.241	0.39	0.98
Venus	.615	0.72	1.01
Earth	1.00	1.00	1.00
Mars	1.88	1.52	1.01
Jupiter	11.8	5.20	0.99
Saturn	29.5	9.54	1.00
Uranus	84.0	19.18	1.00
Neptune	165	30.06	1.00
Pluto	248	39.44	1.00

更多的数据 vs. 更好的算法



互联网需求：大数据处理

- 大数据：规模大、变化快、种类杂

社交类应用

- **Facebook**：用户规模超过10亿，每天新增数据量10TB
- **四大微博(新浪，腾讯、搜狐和网易)**：用户8亿多，每天新增微博超过2亿条，图片2000万张

搜索类应用

- **百度**：每天新增日志数据量近1PB，数据总量近1000PB
- **Google**：每天新处理数据总量已超过20PB

图灵奖得主Jim Gray和IDC报告

- 数据每18月翻一番，过去数据是确定的，当前伴随人机物融合，**网络信息空间大数据呈现多样性和异构性**
- **IDC报告：全球数据2009年0.8 ZB，2012年2.7 ZB，预计2020年达35ZB (2012年的13倍)**

大数据：价值取向？

互联网改变
交流方式

大数据处理改变
经济和社会生活

Google：2007年，通过2万亿单词 训练语言模型，发现简单算法在大数据集时产生更好效果

2008年，通过庞大搜索数据训练4.5亿个数学模型，提前几周预测出H1N1流感的爆发和传播

阿里巴巴：2008年，提前8-9个月预测出金融危机

百度：通过4亿用户分析提供个性化搜索服务

Baidu 百度
Google 谷歌

熟悉用户
浏览行为

淘宝网
Taobao.com
亚马逊
amazon.com

熟悉用户
购物习惯

新浪微博
weibo.com
twitter

了解用户
思维习惯及
社会认知

数据：应用价值？

Twitter：日本海啸、地震信息提前传播，协助紧急事件的应急处理；
微博7.21北京暴雨900万条（受灾分布）、钓鱼岛4000万条（民众情绪）

Google：
2008年在甲型H1N1流感爆发几周前，提前预测冬季流感的传播
阿里巴巴：
提前8-9个月预测08年金融危机；
淘宝网：根据你的消费与浏览商品，判断你可能购买什么。
百度：通过4亿用户分析提供个性化搜索

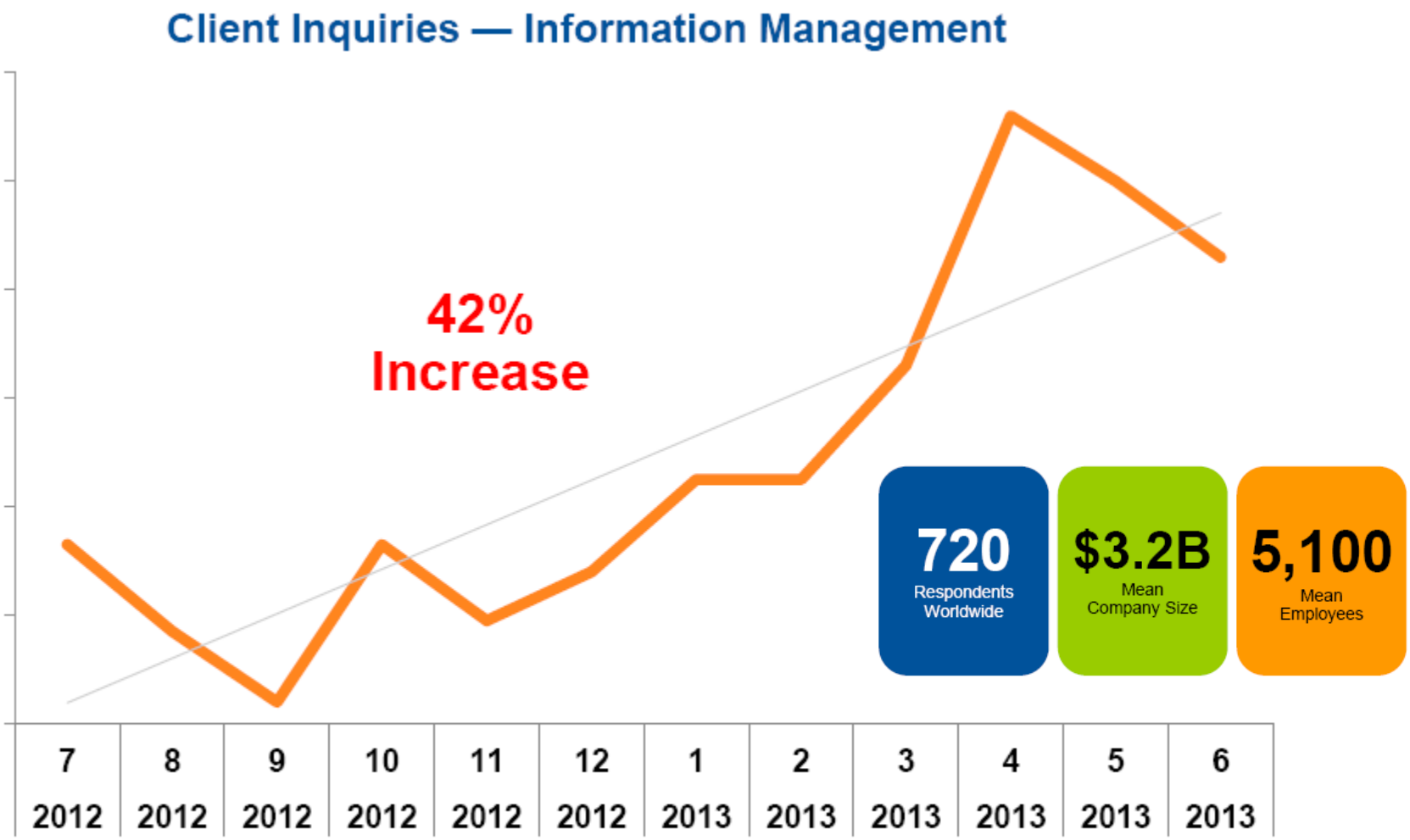
1、根据民众情绪抛售股票。对冲基金从购物网站的顾客评论，分析企业产品销售状况；
2、投资机构搜集分析上市企业声明，寻找破产的蛛丝马迹
3、**根据你关注了哪些人来判断你还可能对哪些人感兴趣。**
美国总统奥巴马的竞选团队依据选民的微博，实时分析选民对总统竞选人的喜好

实质上，通过数据的归类与分析，进行预测，例如出现某种行为的人还很有可能出现另种行为。

提纲

- 大数据定义
- 大数据溯源
- 大数据的应用
- 大数据带来的挑战

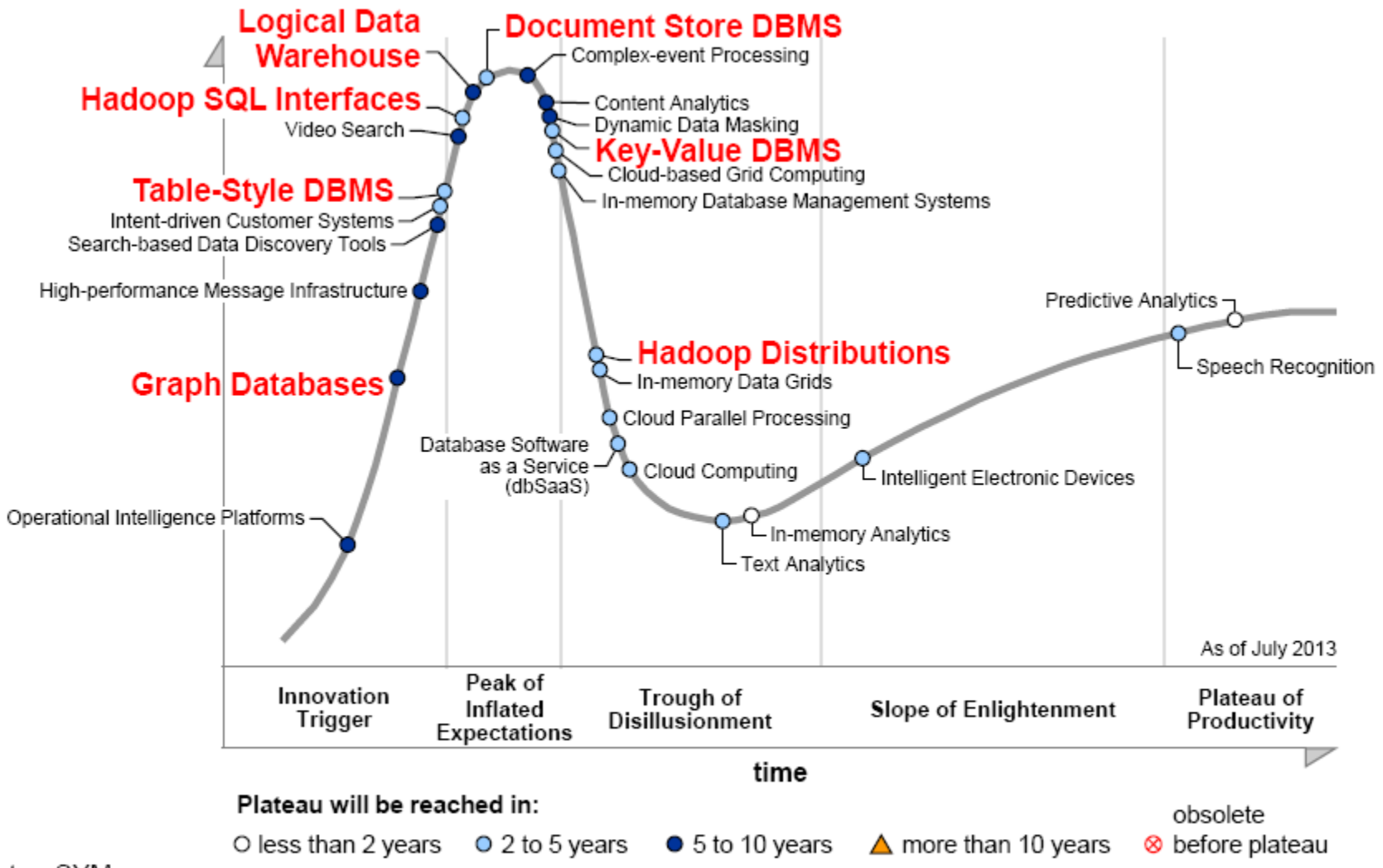
Gartner关于业界对Big Data兴趣的分析



Source: Information Management Team Inquiry Data, July 2012-June 2013

Gartner

Gartner关于Big Data处理技术的分析

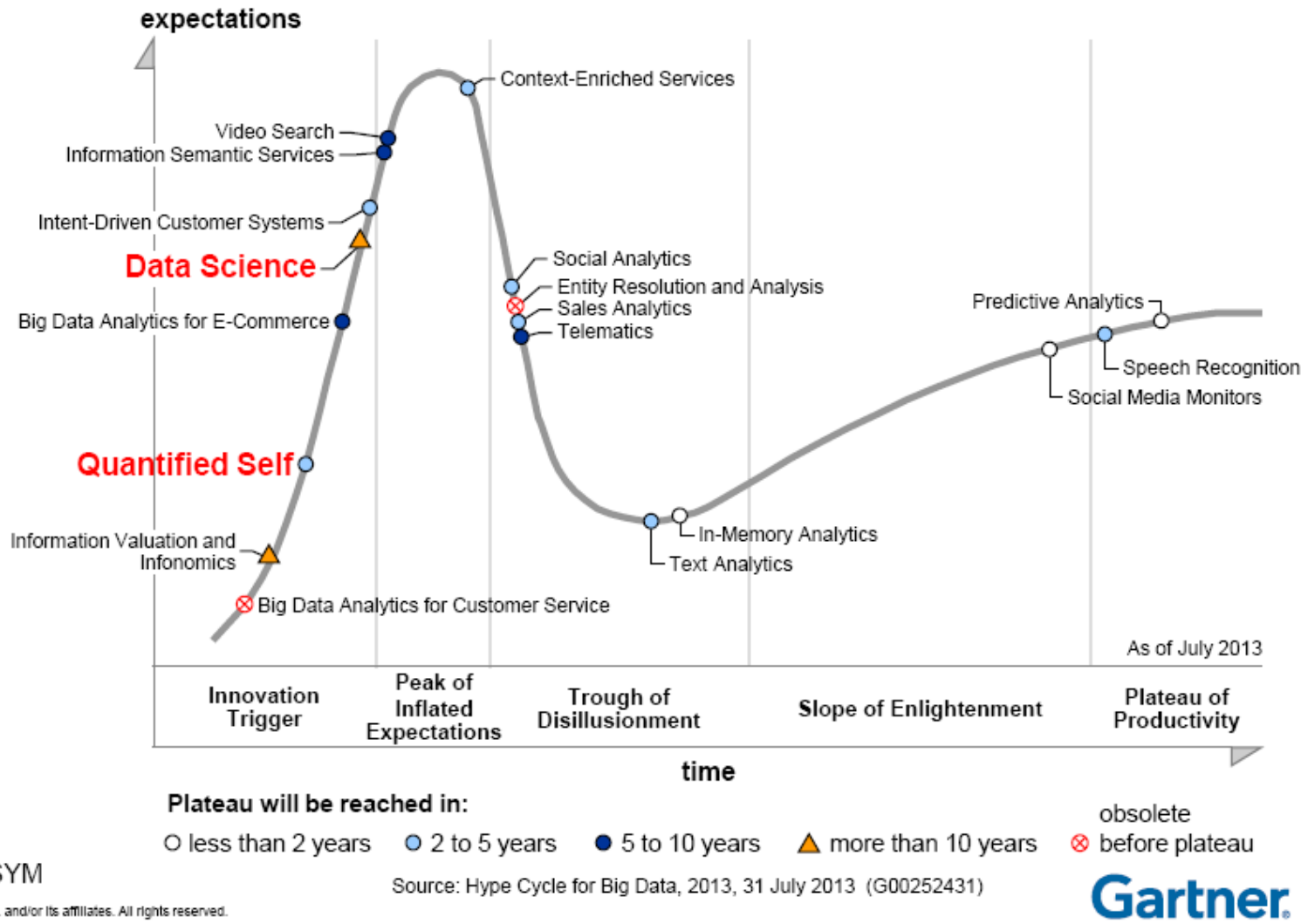


Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

#GartnerSYM

Gartner

Gartner关于Big Data处理技术的分析



#GartnerSYM

Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

Gartner

数据的处理流程

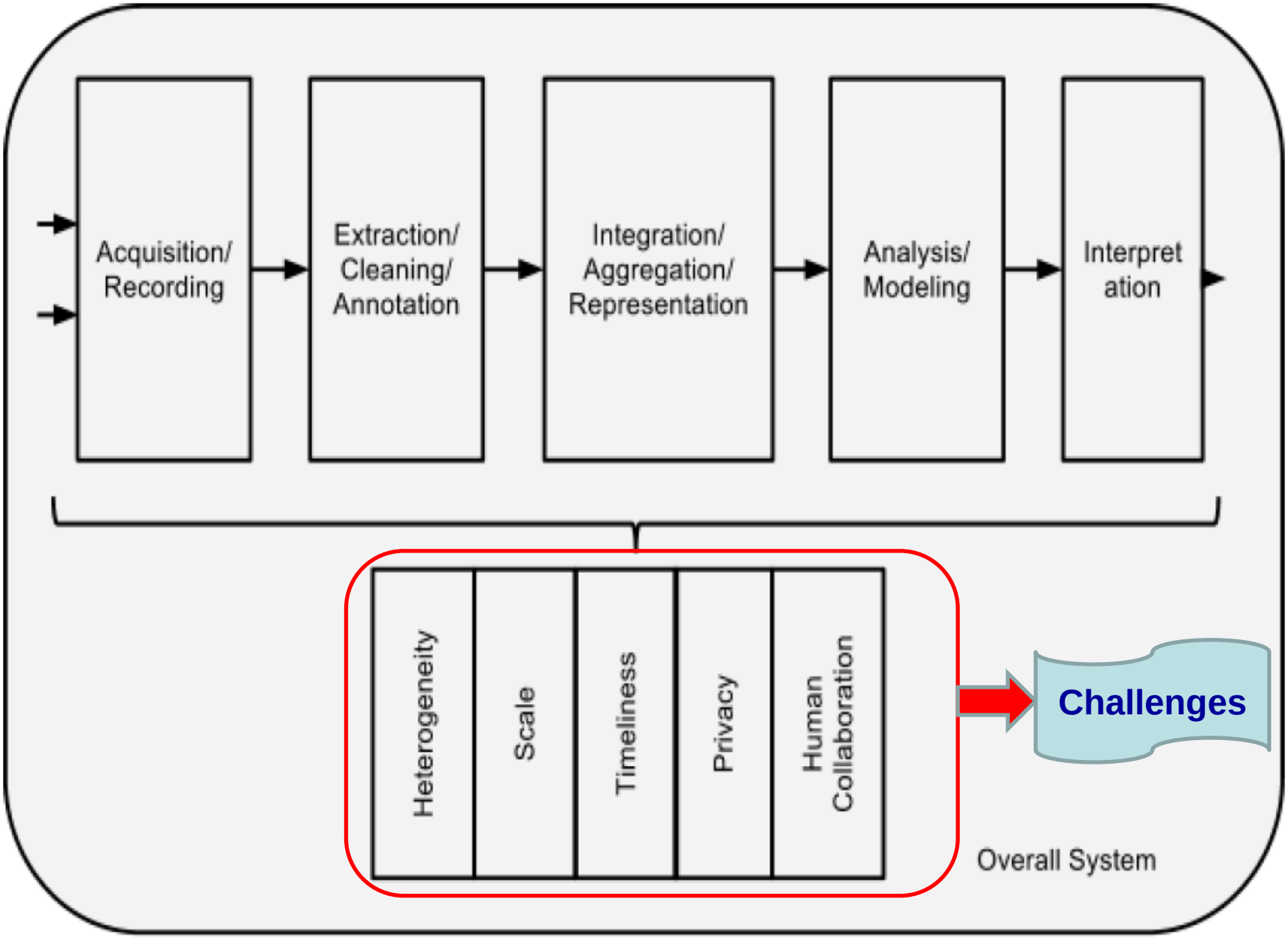
Challenges and Opportunities with Big Data

- **A community white paper** developed by leading researchers across US

Divyakant Agrawal, UC Santa Barbara
Philip Bernstein, Microsoft
Elisa Bertino, Purdue Univ.
Susan Davidson, Univ. of Pennsylvania
Umeshwar Dayal, HP
Michael Franklin, UC Berkeley
Johannes Gehrke, Cornell Univ.
Laura Haas, IBM
Alon Halevy, Google
Jiawei Han, UIUC
Alexandros Labrinidis, Univ. of Pittsburgh

Sam Madden, MIT
Yannis Papakonstantinou, UC San Diego
Jignesh M. Patel, Univ. of Wisconsin
Raghu Ramakrishnan, Yahoo!
Kenneth Ross, Columbia Univ.
Cyrus Shahabi, Univ. of Southern California
Dan Suciu, Univ. of Washington
Shiv Vaithyanathan, IBM
Jennifer Widom, Stanford Univ

A result of remote conversation lasted about 3 months (Nov. 2011 ~ Feb. 2012)

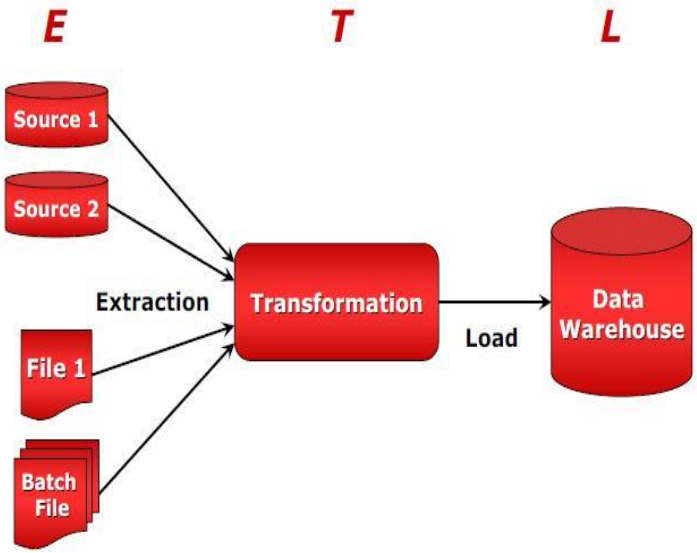

Acquisition/
Recording
Extraction/
Cleaning/
Annotation
Integration/
Aggregation/
Representation
Analysis/
Modeling
Interpret
ation
Heterogeneity
Scale
Timeliness
Privacy
Human
Collaboration
Challenges
Overall System

大数据处理技术分析

- 数据采集
  - ETL工具、爬虫、传感器
- 数据存储
  - 文件系统、关系数据库、图数据库；NoSQL（hadoop）；
- 数据分析
  - NLP、统计、数据挖掘、机器学习、数据库
- 数据展现
- 数据类别
  - 类型（结构、）
  - 行业（医疗、社交）

数据采集–ETL

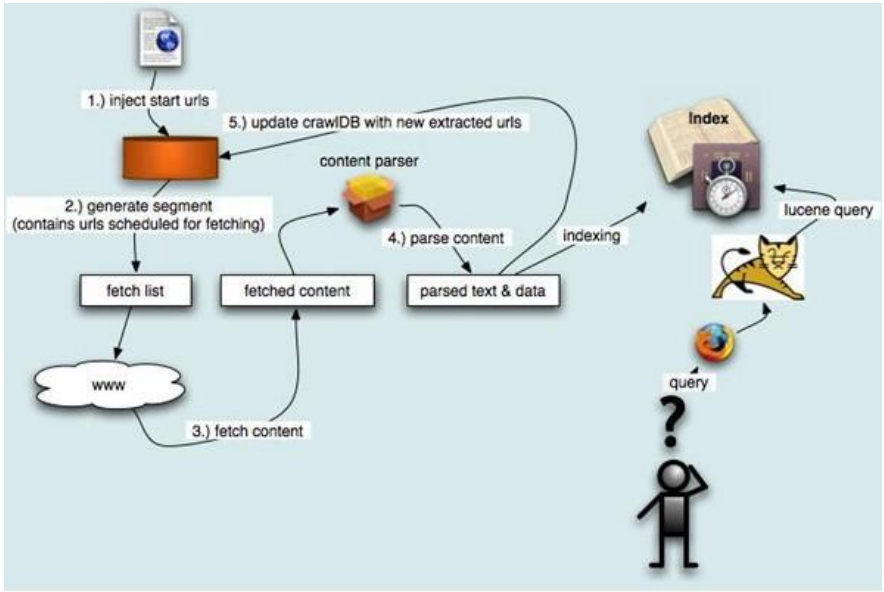
- Extract, Transform and Load (ETL)
  - ETL按照统一的规则集成并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程。



图片来源：http://igorportela.com/extract-transform-and-load-etl/

数据采集-爬虫

- 网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一定停止条件。



图片来源：http://blog.csdn.net/pipi521520/article/details/5599919

数据采集–传感器

- 数据采集是指从传感器和其它待测设备等模拟和数字被测单元中自动采非电量或者电量信号,送到上位机中进行分析，处理。

图片来源http://www.acurite.com/sensor-based-forecasting

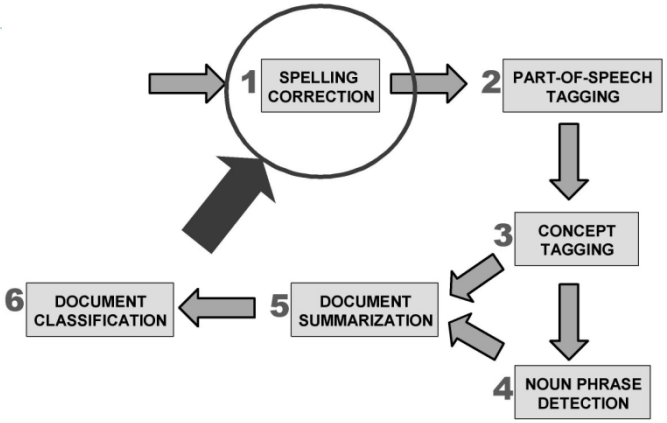
数据存储

- 文件系统
  - 文件数据库又叫嵌入式数据库，将整个数据库的内容保存在单个索引文件中，以便于数据库的发布。
- 关系数据库
  - 关系数据库，是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据
- 图数据库
  - 图数据库的基本含义是以“图”这种数据结构存储和查询数据.
- NoSQL（hadoop）
  - 非关系型数据库以键值对存储（key-value），它的结构不固定，每一个元组可以有不一样的字段，每个元组可以根据需要增加一些自己的键值对，这样就不会局限于固定的结构，可以减少一些时间和空间的开销。

数据处理与分析

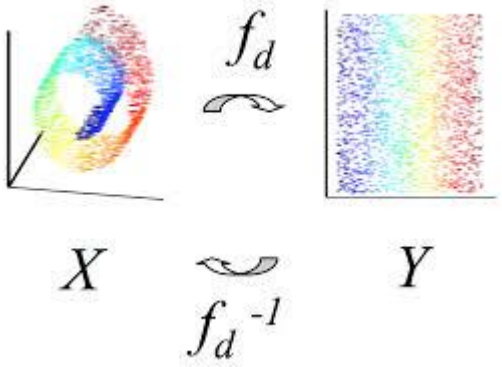
- 数据处理：
  - 自然语言处理技术

实现人与计算机之间用自然语言进行有效通信的各种理论和方法



  - 数据降维技术

将样本点从输入空间通过线性或非线性变换映射到一个低维空间，从而获得一个关于原数据集紧致的低维表示



  - 数据清理技术

发现并纠正*数据*文件中可识别的错误，包括检查*数据*一致性，处理无效值和缺失值等

数据仓库与联机分析处理

- 1988年IBM两位研究人员（Barry Devlin和Paul Murphy）创造性地提出了一个新的术语：数据仓库（Data Warehouse）

- 1992年比尔.恩门出版专著《Building the Data Warehouse》，真正拉开了数据仓库走向大规模应用的序幕，被誉为“数据仓库之父”

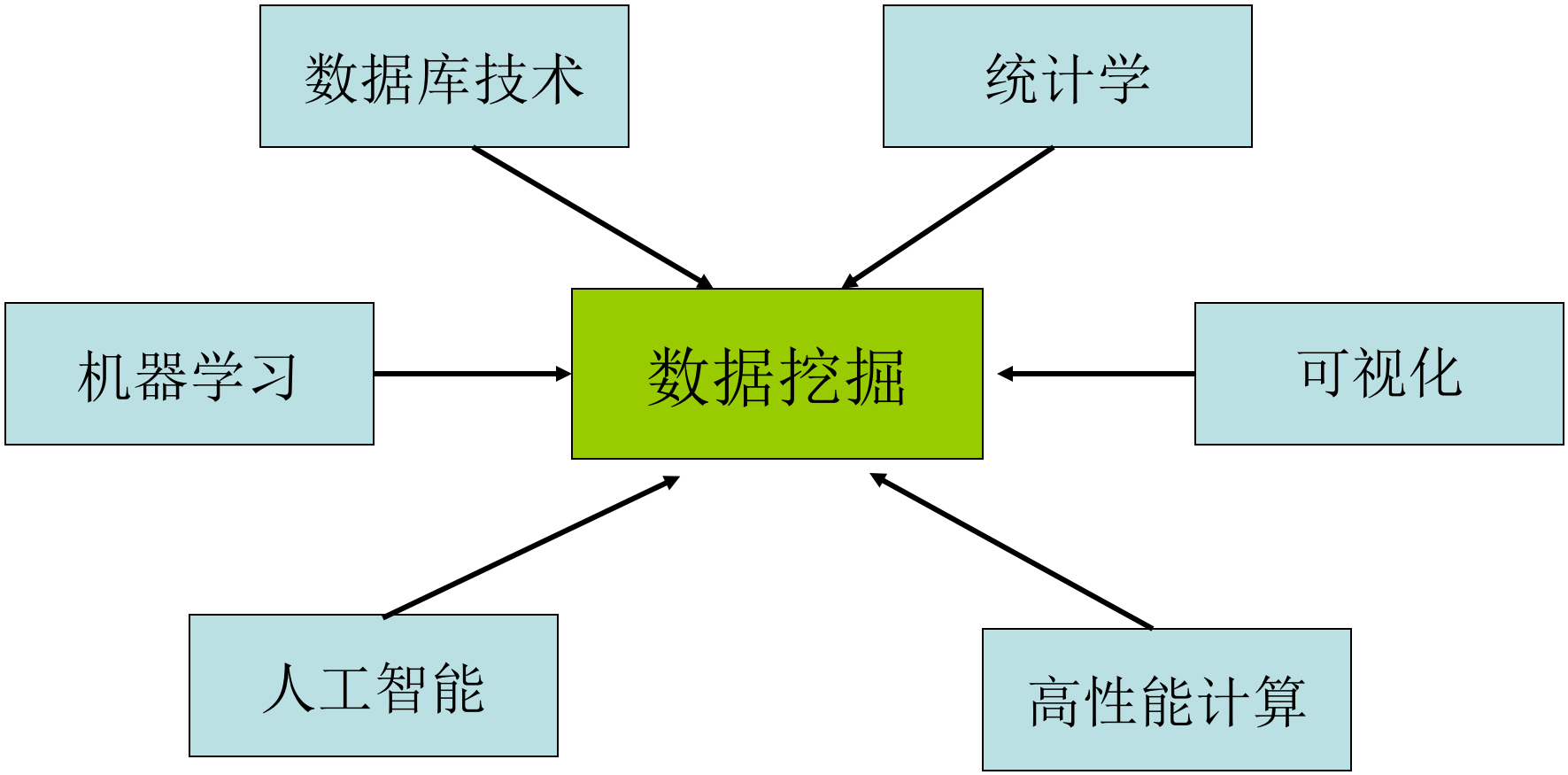
“数据仓库是一个面向主题的、集成的、相对稳定、反映历史变化的数据集合，用于支持管理中的决策制定”

- 数据仓库与数据库的主要区别：
  - 数据仓库以数据分析、决策支持为目的来组织存储数据
  - 数据库的主要目的是为系统保存、查询数据

数据挖掘的定义

- **数据挖掘是从大量数据中提取或“挖掘”知识。**
- 技术上的定义：数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- 商业角度定义：数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。
- 所谓**基于数据库的知识发现**（KDD）是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程。

数据挖掘是多学科的产物



数据挖掘

- **数据挖掘算法按挖掘目的分为：**
  - 关联规则分析
  - 分类与预测
    - ✓ 信息自动分类，信息过滤，图像识别等
  - 聚类分析
  - 异常分析
    - ✓ 入侵检测，金融安全等
  - 趋势、演化分析
    - ✓ 回归，序列模式挖掘

数据挖掘：在你的数据中搜索知识（有趣的模式）

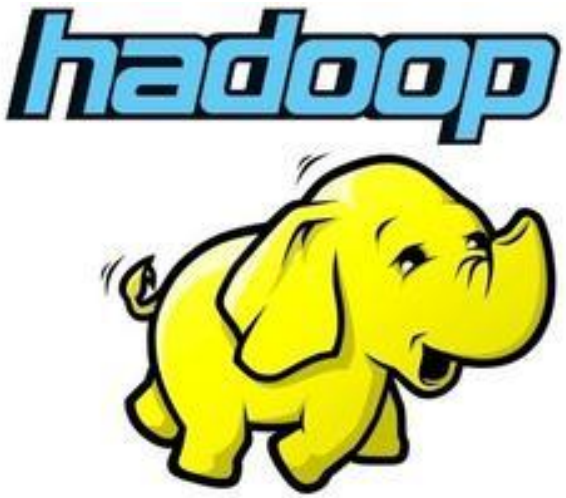
大数据的应用—决策支持

- 1947年，赫伯特. 西蒙在著作《行政组织的决策过程》中指出“人类的理性是有限的，因此所有的决策都是基于有限理论（bounded rationality）的结果”，并指出“如果能利用存储在计算机里的信息来辅助决策，人类理性的范围将会扩大，决策的质量就能提高”
- 预测“在后工业时代，也就是信息时代，人类社会面临的中心问题将从如何提高生产庇转变为如何更好地利用信息来辅助决策”

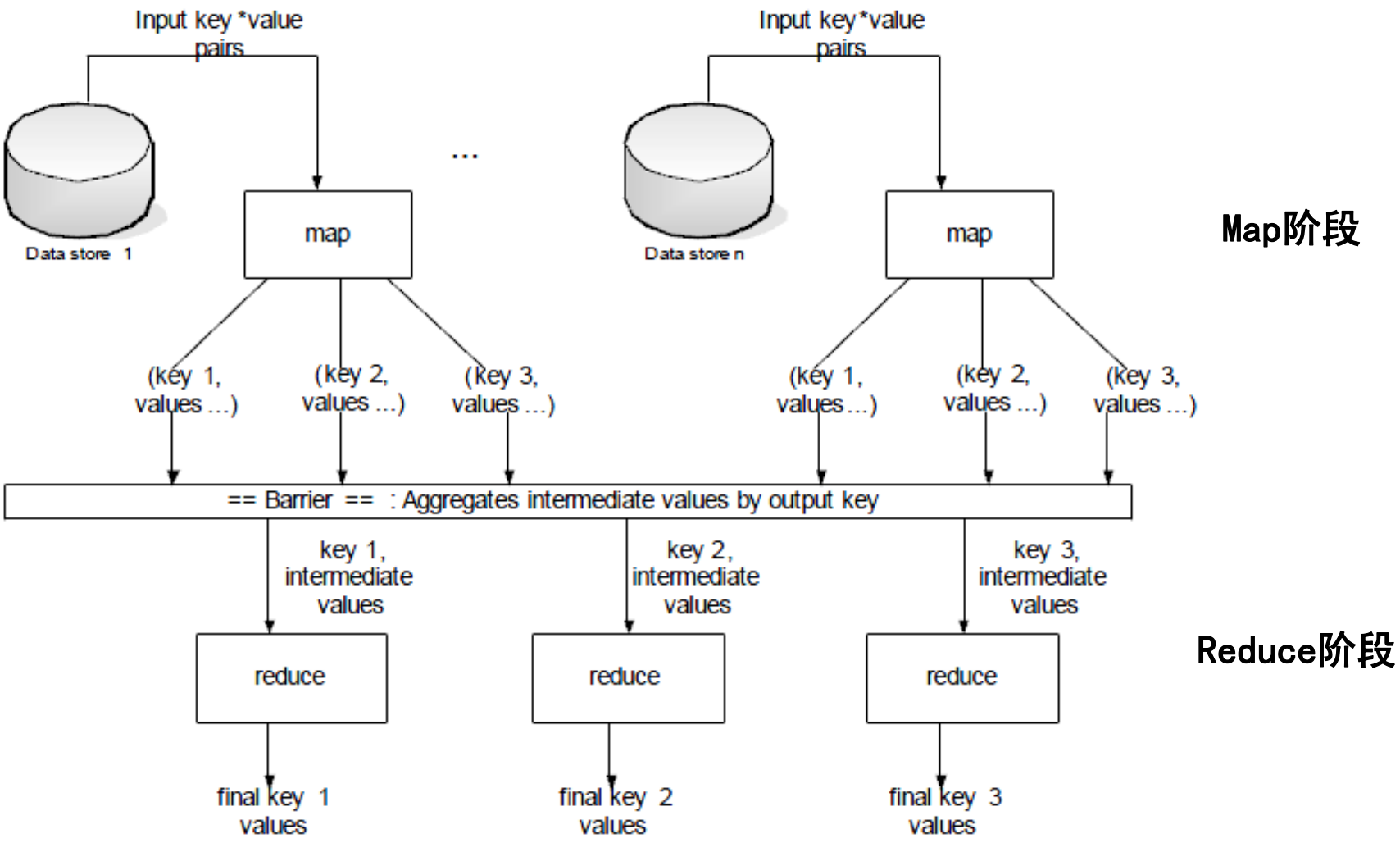
1975年图灵奖
1978年诺贝尔经济学奖
1993年美国心理协会终身成就奖

MapReduce/Hadoop and Beyond

- 由Google提出的一个用于大数据处理的系统
  - Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, OSDI 2004.
- Apache开源社会项目： Hadoop
- 主要的思想来自于functional programming



MapReduce/Hadoop and Beyond



MapReduce/Hadoop and Beyond

- MapReduce/Hadoop的局限性
  - 比较底层的编程模型
  - 对实时处理和递归处理支持不够
  - 适合处理具有“局部性”的数理

- Beyond MapReduce
  - 高层编程语言：Hive (Facebook), Pig (Yahoo!)等…
  - 流式计算：S4 (Yahoo!), Storm (Twitter), Spark (UC Berkeley AMP lab)
  - 支持递归的系统：Google Pregel
  - 其他技术。。。

2004

- Jeffrey Dean, Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters. OSDI 2004: 137-150

2008

- David J. DeWitt and Michael Stonebraker, MapReduce: A major step backwards, The Database Column, January 17, 2008 4:20 PM

① A giant step backward in the programming paradigm for large-scale data intensive applications
② A sub-optimal implementation, in that it uses brute force instead of indexing
③ Not novel at all -- it represents a specific implementation of well known techniques developed nearly 25 years ago
④ Missing most of the features that are routinely included in current DBMS
⑤ Incompatible with all of the tools DBMS users have come to depend on

2009

- Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, Michael Stonebraker: A comparison of approaches to large-scale data analysis. SIGMOD, 2009.

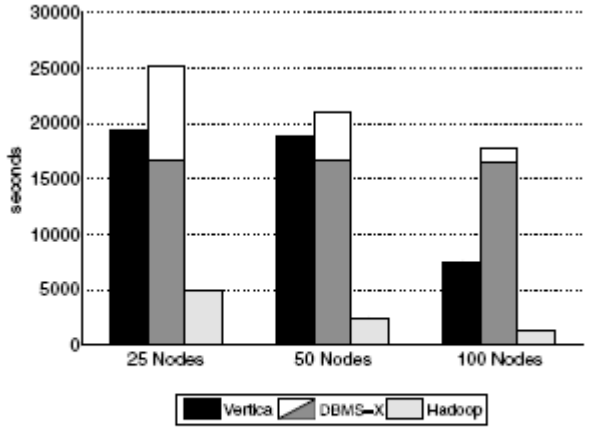


Figure 2: Load Times – Grep Task Data Set (1TB/cluster)

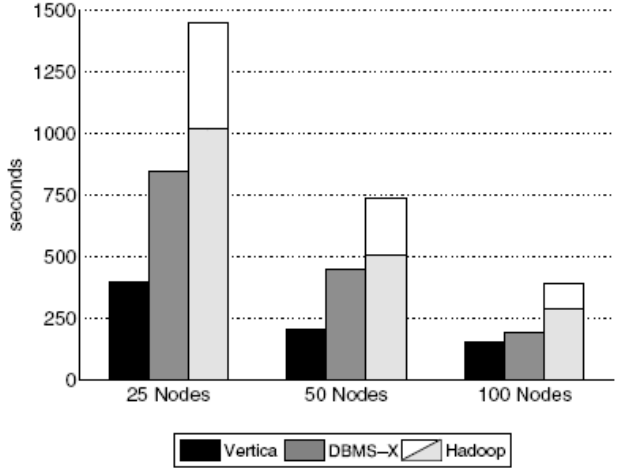


Figure 5: Grep Task Results – 1TB/cluster Data Set

"Grep task" taken from the original MapReduce paper

2009

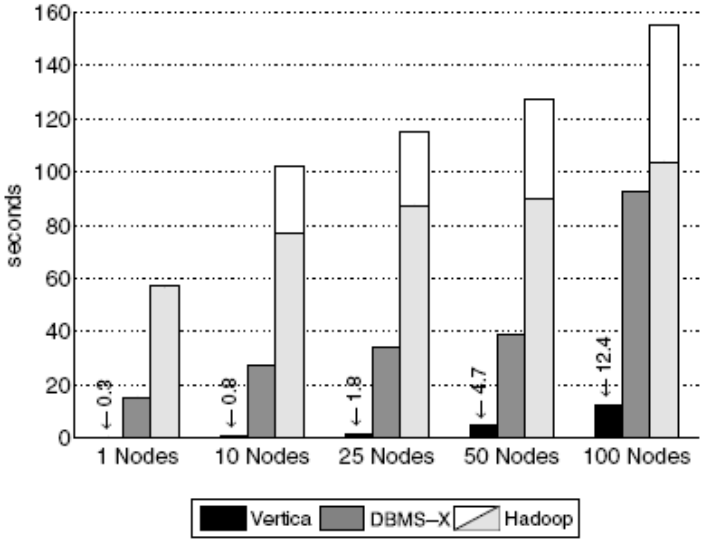


Figure 6: Selection Task Results

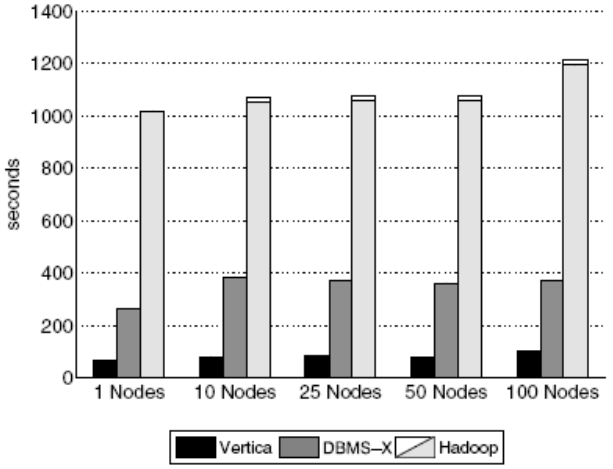


Figure 8: Aggregation Task Results (2,000 Groups)

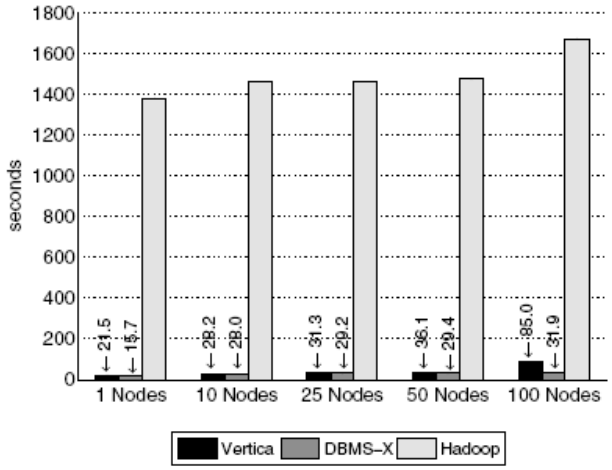


Figure 9: Join Task Results

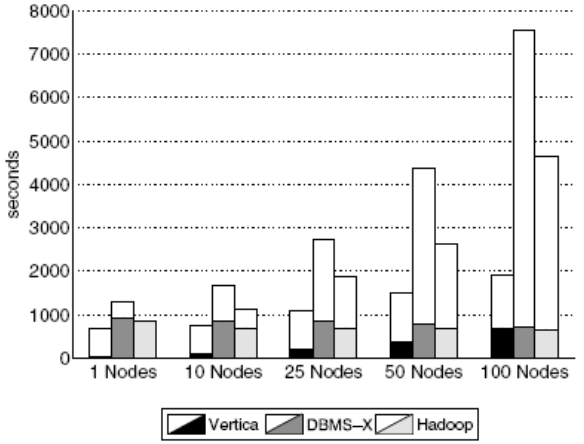


Figure 10: UDF Aggregation Task Results

2010

- Dawei Jiang, Beng Chin Ooi, Lei Shi, Sai Wu: The Performance of MapReduce: An In-depth Study. PVLDB 3(1): 472-483 (2010)
-

Table 1: Indirect Comparison with Parallel Database Systems

	DBMS-X	Vertica	HadoopOpt
Grep	1.5x	2.6x	1.47x
Aggregation (Large)	1.6x	4.3x	1.54x
Join	36.3x	21.0x	14.68x

2010

- Communications of the ACM, Vol. 53 No. 1
- Michael Stonebraker, Daniel Abadi, David J. DeWitt, Sam Madden, Erik Paulson, Andrew Pavlo, Alexander Rasin,MapReduce and Parallel DBMSs: Friends or Foes?
- Jeffrey Dean, Sanjay Ghemawat, MapReduce: A Flexible Data Processing Tool

	RDBMS	MapReduce
模式	内部支持	外部附加
索引	内部支持	编程实现
数据类型	结构化数据	非结构化、半结构化、结构化数据
编程模型	声明性语言 SQL	过程性语言
灵活性	有限	大
扩展性	上百节点	上千节点
容错性	低,查询重启	高,子任务重新执行
性能	高	比 RDBMS 低 [1]
应用范围	在线事务处理 在线分析处理	批量处理 [2] 深度分析

2014

- Spark vs. MapReduce, http://databricks.com/

	Hadoop保持记录	Spark 100 TB	Spark 1 PB
数据大小	102.5 TB	102 TB	1000 TB
耗时	72分钟	23分钟	234分钟
节点数	2100	206	190
# Cores	50400	6592	6080
# Reducers	10,000	29,000	250,000
Rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min
Daytona Gray类别排序基准规则	是	是	否
环境	专用的数据中心	EC2 (i2.8xlarge)	EC2 (i2.8xlarge)

大数据可视化

- 数据可视化
  - 主要旨在借助于图形化手段，清晰有效地传达与沟通信息。
  - 美学形式与功能齐头并进；通过直观地传达关键的方面与特征，实现对于相当稀疏而又复杂的数据集的深入洞察。
- 数据可视化的分类 （ Frits H. Post, Gregory M. Nielson and Georges-Pierre Bonneau (2002). *Data Visualization: The State of the Art*. ）
  - 可视化算法与技术方法
  - 立体可视化
  - 信息可视化
  - 多分辨率方法
  - 建模技术方法
  - 交互技术方法与体系架构

核医学成像

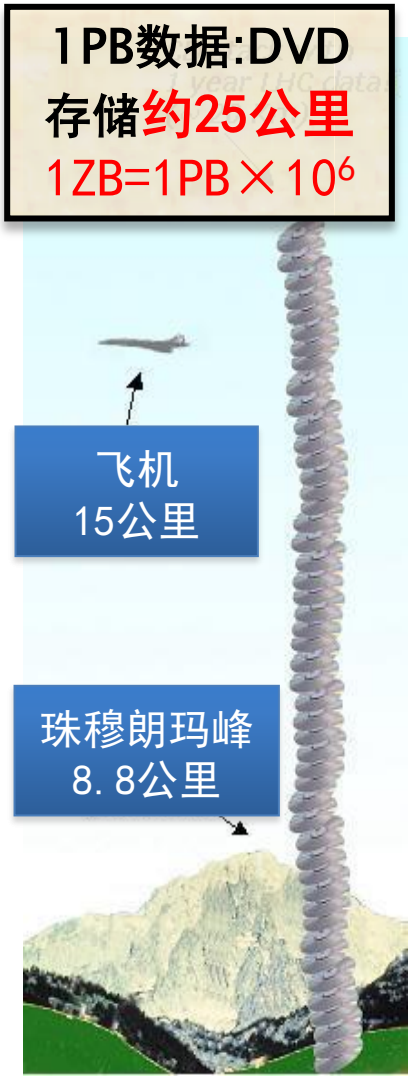
螺旋星云可见光图像

大数据类别

- 数据类型
  - 结构化数据：
    - 关系数据等：数据的查询、统计、更新等操作效率低。
  - 半结构化数据：
    - XML、图数据等：转换为结构化存储或者按照非结构化存储。
  - 非结构化数据：
    - 图片、视频、word、pdf、ppt等：不利于检索、查询和存储
- 行业数据
  - 大规模的电子商务数据
  - 社会数据（社会网络，互联网等），是一类重要的图数据
  - 移动数据(呼叫详细记录、 RFID、传感器网络)
  - 医疗数据
  - 天文学，大气科学，基因组学，生物地球化学，生物和其他复杂和/或跨学科的科研数据

大数据现状

- **Google**：通过大规模集群和 MapReduce 软件，每个月处理 的数据量超过 400PB。
- **百度**：数百 PB，每天大约要处理几十 PB 数据，大多 要实时处理，如微博、团购、秒杀。
- **Facebook**：注册用户超过 10亿，每月上传 10 亿照片，每天生 成 300TB 日志数据
- **淘宝**：有 3.7 亿会员，在线商品 8.8 亿，每天交易数千万， 产生约 20TB 数据。
- **Yahoo!**：Hadoop 云计算平台有 34 个集群，超过 3 万 台机器，总存储容量超过 100PB。



医疗大数据

- 据统计，目前我国重大病患者有近2.6亿人，同时老龄化严重：60岁以上的老年人已经达2.02亿，此外，全国还有大概8千万的残疾人，这都是我国目前面临的一个严重的健康事业难题 - **医疗器械创新网**www.innomd.org
- 现在人类已知的疾病，大概有1万种，各国批准的临床诊断标准，有标准的诊断方法，全球批准了大概3000种。美国人批准的药物是4600种，粗粗的算了一下中国是3000种，中国有3万-4万家医院，近1千万的医务人员，这几个数字列在一起，3000种，3-4万，1千万，**最后的结论就是三个字，就是不靠谱，到医院去不靠谱。-华大基因董事长汪建**
- **Big data redefines the traditional scientific methods used in medicine**(www.techrepublic.com)
  - 斯坦福大学将于2015年5月20到22日举办一个生物医学领域的大数据会议，该会议针对各大高校、医院、政府部门和机构的医学研究人员，旨在鼓励合作、应对挑战以及建立在医疗保健领域使用大数据的可行步骤。
  - 会议通告中写到：“在从大数据的大规模整合及分析中攫取价值这方面，其它行业已经取得了极大成功，而医疗健康行业才刚起步（沾湿了脚丫，getting its feetwet）。**是的，医疗健康的提供者（如医疗机构等）和付费者（如病人等）正日益增加在分析能力上的投入，以更好地理解不断变化的健康医疗环境，但这还只是处于初级阶段。”**

医疗大数据

- **医生们需要数据**
  - 2011年的斯坦福Lucile Packard儿童医院，一位来自内华达州里诺的女孩被用直升机送到该医院的加护病房（ICU）。她患有狼疮，一种攻击人体健康组织并能导致永久性肾损伤的疾病。一个多学科医生团队不得不在使用凝结剂和复合手术的风险间权衡，凝结剂能够稀释血液以防止血液结块，复合手术会导致中风或者器官内出血。
- **基于数据的医疗手段**
  - 一位叫Jennifer Frankovich的年青医师诉诸于使用狼疮患儿数据库，她曾参与建立该数据库。作为数据库工作的一部分，需要将图表数字化并使数据可通过关键字来检 索。通过搜索数据库，Frankovich医生能够查阅每位来院的狼疮患儿，从而了解他们中出现血凝现象的人数，以及导致危险的因素。据此，她可以计算出 使用抗凝血剂的风险能否佐证小女孩出现血液凝块的风险。计算结果表明，值得冒这个险：使用了抗凝血剂后，小女孩的病情出现了好转迹象。
- **需要解析数据的系统**
  - 医院的管理层仍然认为，对于紧急病例，相比于查找过往成功案例的医疗数据，相信医师团队的集体智慧更加安全稳妥。在今年一月份接受NPR采访 时，Frankovich医生坦言，“分析数据是一个复杂的工作，需要特定的专业知识和技能。试想，假若搜索引擎有程序错误，亦或档案被错误的转录，后果将会如何？真的有太多地方会出现错误… …。这将需要一个系统来解析数据，而这样的系统是我们还尚未拥有。”
- **观点：确实如此，但是在诸如克利夫兰临床中心（ClevelandClinic）这样的医疗健康机构中，医生和医学实践者们已经在利用大数据和分析 法来诊断病情和实施治疗。当跨学科医生团队评估病人时，数据分析结果已然进入了他们的讨论之中。并且，尽管医疗健康数据的质量和整合问题将持续存在，无容 置疑的是，重新定义传统科学方法已初现端倪。**

医疗大数据与互联网

德国	工业4.0	**2013年4月在汉诺威工业博览会(Cebit)正式提出，上升为国家级战略** 将生产设备联网，灵活智能地配置**生产要素**，将制造业向智能化转型
美国	工业互联网	**美国通用电气（GE）最早提出概念，由五家行业龙头（GE、Intel、IBM、Cisco、AT&T）组建工业互联网联盟(IIC)** 通过网络、传感器等技术，实现机器间的连接并且最终将人机连接，结合**软件和大数据分析**，重构全球工业
中国	互联网+	**2015年 李克强总理在政府工作报告中首次提出“互联网+”行动计划，上升为国家级战略** 充分发挥互联网在生产要素配置中的优化和集成作用，将互联网的创新成果深度融合于经济社会各领域之中，提高**实体经济**的创新力和生产力，形成更广泛的以互联网为基础设施和实现工具的经济发展新形态

来自中国移动研究院首席科学家**许利群**

医疗服务机构纷纷拥抱移动互联网

医院诊所

- **北京大学人民医院** 2010年开展“**实景医学**”**模式**研究，即主要利用无线互联网技术，使患者在日常生活工作状态下，动态实时接收医疗机构的某些诊断、监护、治疗的医疗服务，在医院外建立了一个新的全天候诊疗服务平台。
- **Scripps Clinic** Dr. Eric Topol 执业心脏病专家，基因组学教授写的两本具有广泛影响的书。
- **Mayo Clinics**为患者提供个人健康助理、护士热线、自动症状检查和电子病历等多种移动健康服务，并提供专业的糖尿病管理应用。

MAYO CLINIC

体检机构

- **慈铭体检**推出O2O运动健康管理产品。包含了智能运动腕表、慈铭O2O健康管理手机APP以及三位一体的健康管理云平台，提供体检数据存储、远程医疗咨询、健康行为干预等。
- **爱康国宾**推出360度健康管理app，提供体检预约、购买等功能，此外用户可随时通过手机客户端与爱康国宾进行咨询、互动。

连锁药店

- **Walgreens** 是美国最大的零售药店，拥有8200家连锁店，其健康选择（healthy choices）项目Balance Rewards计划**成员达到8200万**。这些成员使用移动健康设备和应用跟踪健康状态，并且获得积分奖励。

在Walgreens的**移动APP**中，由经过认证的专业医生可为用户提供24小时的虚拟随访服务。另外接入病友社区**PatientsLikeMe**中关于药物副作用的数据。

Walgreens
The Pharmacy America Trusts

来自中国移动研究院首席科学家**许利群**

阿里未来医院计划：以支付能力切入医疗领域

“未来医院计划”一阶段

- **帮助医院建立移动医疗的服务体系。**支付宝与医院的HIS系统相对接，每个人的就诊卡绑定支付宝账户，从而实现病人在单个医院内部信息的整合、各个医疗机构间（医院、疾控中心、诊所、体检医院、药店）的信息整合，形成“阿里中国医疗健康云”。
- 医院入驻支付宝钱包的“服务窗”之后，用户就可以通过支付宝钱包完成挂号、远程候诊、诊间缴费、取报告单、诊后互动等多个就医环节，而不需要多次排队等待。
- 目前阿里支付宝已与上海长海医院、上海第一妇婴保健院等上海、杭州、厦门和北京等城市的多个医院进行了对接；与全国主要城市的近50家三甲医院达成合作意向。

“未来医院计划”二阶段

- 将结合医疗改革的推进，通过互联网在线完成**电子处方**、就近药物配送、**转诊**、医保实时报销、商业保险实时申赔等所有环节。**2014年12月，阿里健康APP的上线**，标志着这一计划进入商用阶段。

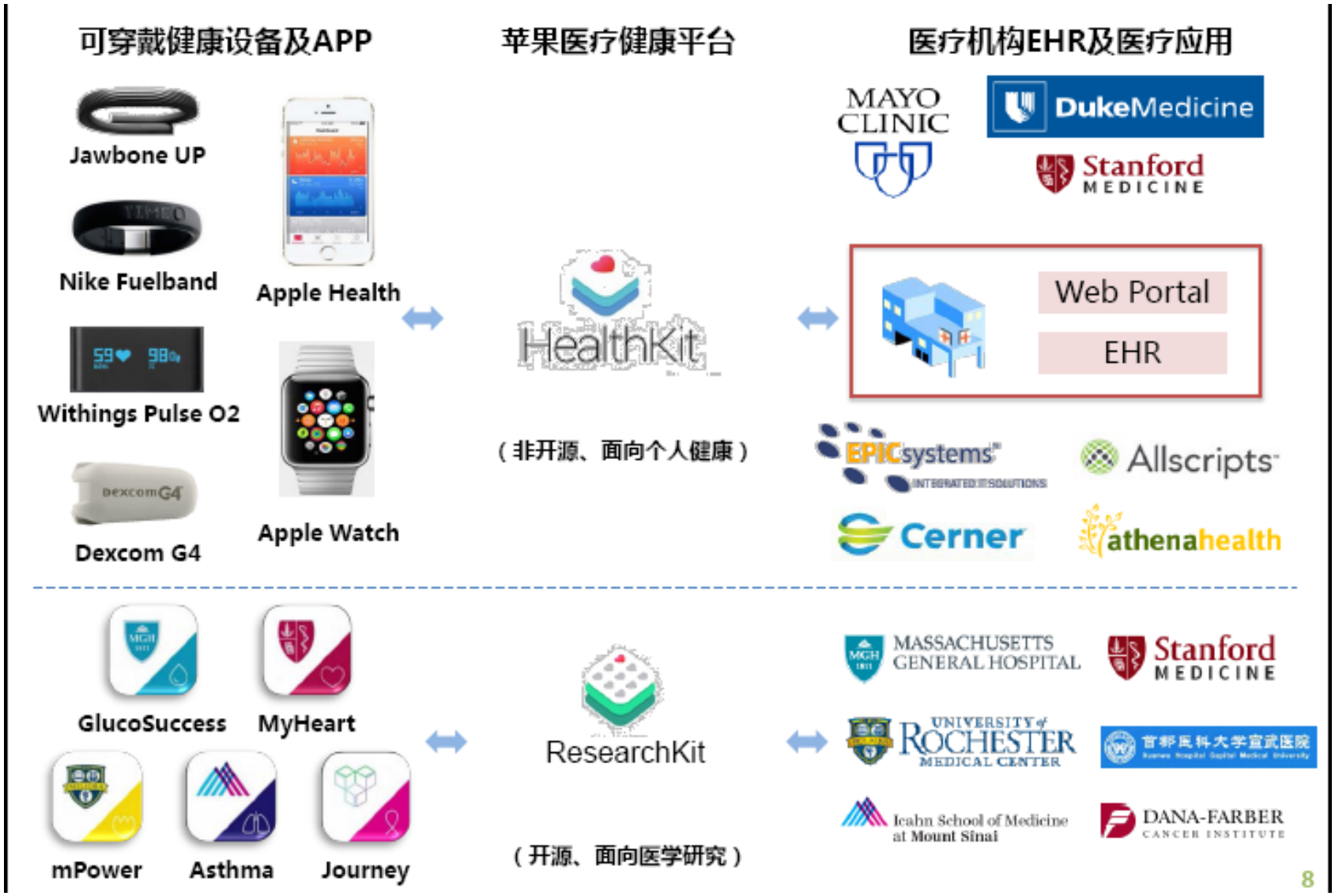
“未来医院计划”三阶段

- 支付宝将开放大数据平台，结合云计算能力，与可穿戴设备厂商、医疗机构、政府卫生部门等合作，共同搭建基于大数据的健康管理平台，**实现从治疗到预防的转变。**

投资布局

来自中国移动研究院首席科学家**许利群**

苹果：全方位布局医疗健康



来自中国移动研究院首席科学家**许利群**

小结

- 大数据是产业＋资源＋科学，其发展环境支撑互联网科技创新、产业应用发展和政策保障，形成“生态链”
- 需要智者：多学科融合
  - 教育思考＋科研人员＋产业领袖
- 需要实践者：真实的数据和计算平台
  - 开放的数据服务与共享平台
  - 产学研紧密合作
- 需要政策：支撑和政府支持


谢谢