课程网址：http://dblab.xmu.edu.cn/node/422

=课程教材由林子雨老师根据网络资料编著=

厦门大学计算机科学系教师 林子雨 编著

http://www.cs.xmu.edu.cn/linziyu

2013 年 9 月

前言

本教程由厦门大学计算机科学系教师林子雨编著，可以作为计算机专业研究生课程《大数据技术基础》的辅助教材。

本教程的主要内容包括：大数据概述、大数据处理模型、大数据关键技术、大数据时代面临的新挑战、NoSQL 数据库、云数据库、Google Spanner、Hadoop、HDFS、HBase、MapReduce、Zookeeper、流计算、图计算和 Google Dremel 等。

本教程是林子雨通过大量阅读、收集、整理后，精心制作的学习材料，与广大数据库爱好者共享。教程中的内容大部分来自网络资料和书籍，一部分是自己撰写。对于自写内容，林子雨老师拥有著作权。

本教程 PDF 文档及其全套教学 PPT 可以通过网络免费下载和使用（下载地址：http://dblab.xmu.edu.cn/node/422）。教程中可能存在一些问题，欢迎读者提出宝贵意见和建议！

本教程已经应用于厦门大学计算机科学系研究生课程《大数据技术基础》，欢迎访问 2013 班级网站 http://dblab.xmu.edu.cn/node/423。

林子雨的 E-mail 是：ziyulin@xmu.edu.cn。

林子雨的个人主页是：http://www.cs.xmu.edu.cn/linziyu。

林子雨于厦门大学海韵园

2013 年 9 月

第 1 章 大数据概述

厦门大学计算机科学系教师 林子雨 编著

http://www.cs.xmu.edu.cn/linziyu

2013 年 9 月

第 1 章 大数据概述

大数据时代已经到来。最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡，麦肯锡称：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”“大数据”在物理学、生物学、环境生态学等领域以及军事、金融、通讯等行业存在已有时日，却因为近年来互联网和信息行业的发展而引起人们关注。

本章内容旨在让大家更好地认识和了解大数据，要点如下：

- 大数据概念
- 大数据的产生和应用
- 大数据作用
- 大数据与大规模数据、海量数据的差别
- 典型的大数据应用实例
- 从数据库到大数据
- 大数据与云计算
- 大数据与物联网
- 对大数据的错误认识
- 大数据技术
- 大数据存储和管理技术
- 大数据生态系统

1.1 大数据概念

随着以博客、社交网络、基于位置的服务 LBS 为代表的新型信息发布方式的不断涌现，以及云计算、物联网等技术的兴起，数据正以前所未有的速度在不断的增长和累积，大数据时代已经来到。

根据 IDC 作出的估测，数据一直都在以每年 50%的速度增长，也就是说每两年就增长一

倍（大数据摩尔定律）。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量，预计到 2020 年，全球将总共拥有 35ZB 的数据量，相较于 2010 年，数据量将增长近 30 倍。这不是简单的数据增多的问题，而是全新的问题。举例来说，在当今全球范围内的工业设备、汽车、电子仪表和装运箱中，都有着无数的数字传感器，这些传感器能测量和交流位置、运动、震动、温度和湿度等数据，甚至还能测量空气中的化学变化。将这些交流传感器与计算智能连接起来，就是目前“物联网”(Internet of Things)或“工业互联网”(Industrial Internet)。在信息获取的问题上取得进步是促进“大数据”趋势发展的重要原因。物联网、云计算、移动互联网、车联网、手机、平板电脑、PC 以及遍布地球各个角落的各种各样的传感器，无一不是数据来源或者承载的方式。

学术界、工业界甚至于政府机构都已经开始密切关注大数据问题，并对其产生浓厚的兴趣。就学术界而言，Nature 早在 2008 年就推出了 Big Data 专刊。计算社区联盟(Computing Community Consortium)在 2008 年发表了报告《Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society》，阐述了在数据驱动的研究背景下，解决大数据问题所需的技术以及面临的一些挑战。Science 在 2011 年 2 月推出专刊《Dealing with Data》，主要围绕着科学研究中大数据的问题展开讨论，说明大数据对于科学研究的重要性。美国一些知名的数据管理领域的专家学者则从专业的研究角度出发，联合发布了一份白皮书《Challenges and Opportunities with Big Data》。该白皮书从学术的角度出发，介绍了大数据的产生，分析了大数据的处理流程，并提出大数据所面临的若干挑战。

全球知名的咨询公司麦肯锡(McKinsey)去年 6 月份发布了一份关于大数据的详尽报告《Big data: The next frontier for innovation, competition, and productivity》，对大数据的影响、关键技术和应用领域等都进行了详尽的分析。进入 2012 年以来，大数据的关注度与日俱增。1 月份的达沃斯世界经济论坛上，大数据是主题之一，该次会议还特别针对大数据发布了报告《Big Data, Big Impact: New Possibilities for International Development》，探讨了新的数据产生方式下，如何更好的利用数据来产生良好的社会效益。该报告重点关注了个人产生的移动数据与其他数据的融合与利用。3 月份美国奥巴马政府发布了《大数据研究和发展倡议》(Big Data Research and Development Initiative)，投资 2 亿以上美元，正式启动“大数据发展计划”。计划在科学研究、环境、生物医学等领域利用大数据技术进行突破。奥巴马政府的这一计划被视为美国政府继信息高速公路(Information Highway)计划之后在信息科学领域的又一重大举措。与此同时，联合国一个名为 Global Pulse 的倡议项目在今年 5 月发布报告《Big Data for Development：Challenges & Opportunities》，该报告主要阐述大数据时代各国特别是发展中

国家在面临数据洪流(Data Deluge)的情况下所遇到的机遇与挑战，同时还对大数据的应用进行了初步的解读。《纽约时报》的文章《The Age of Big Data》则通过主流媒体的宣传使普通民众开始意识到大数据的存在，以及大数据对于人们日常生活的影响。

可以看出，“大数据”是时下最火热的 IT 行业词汇。其实，早在 1980 年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。不过，大约从 2009 年开始，“大数据”才成为互联网信息技术行业的流行词汇。

那么，如何给大数据下一个定义呢？一般而言，大家比较认可关于大数据的 4V 说法。大数据的 4 个“V”，或者说是大数据的四个特点，包含四个层面：第一，数据体量巨大。从 TB 级别，跃升到 PB 级别；第二，数据类型繁多。前文提到的网络日志、视频、图片、地理位置信息等等。第三，价值密度低，商业价值高。以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒。第四，处理速度快。1 秒定律。最后这一点也是和传统的数据挖掘技术有着本质的不同。业界将其归纳为 4 个“V”——Volume，Variety，Value，Velocity。

舍恩伯格的《大数据时代》受到了广泛的赞誉，他本人也因此书被视为大数据领域中的领军人物。在舍恩伯格看来，大数据一共具有三个特征：（1）全样而非抽样；（2）效率而非精确；（3）相关而非因果。

第一个特征非常好理解。在过去，由于缺乏获取全体样本的手段，人们发明了“随机调研数据”的方法。理论上，抽取样本越随机，就越能代表整体样本。但问题是获取一个随机样本代价极高，而且很费时。人口调查就是典型一例，一个稍大一点的国家甚至做不到每年都发布一次人口调查，因为随机调研实在是太耗时耗力了。

但有了云计算和数据库以后，获取足够大的样本数据乃至全体数据，就变得非常容易了。谷歌可以提供谷歌流感趋势的原因就在于它几乎覆盖了 7 成以上的北美搜索市场，而在这些数据中，已经完全没有必要去抽样调查这些数据：数据仓库，所有的记录都在那里躺着等待人们的挖掘和分析。

第二点其实建立在第一点的基础上。过去使用抽样的方法，就需要在具体运算上非常精确，因为所谓“差之毫厘便失之千里”。设想一下，在一个总样本为 1 亿人口随机抽取 1000 人，如果在 1000 人上的运算出现错误的话，那么放大到 1 亿中会有多大的偏差。但全样本时，有多少偏差就是多少偏差而不会被放大。谷歌人工智能专家诺维格，在他的论文中写道：大数据基础上的简单算法比小数据基础上的复杂算法更加有效。

数据分析的目的并非仅仅就是数据分析，而是有其它用途，故而时效性也非常重要。精确的计算是以时间消耗为代价的，但在小数据时代，追求精确是为了避免放大的偏差而不得

已为之。但在样本=总体的大数据时代，“快速获得一个大概的轮廓和发展脉络，就要比严格的精确性要重要得多”。

第三个特征则非常有趣。相关性表明变量 A 和变量 B 有关，或者说 A 变量的变化和 B 变量的变化之间存在一定的正比（或反比）关系。但相关性并不一定是因果关系（A 未必是 B 的因）。

亚马逊的推荐算法非常有名，它能够根据消费记录来告诉用户你可能会喜欢什么，这些消费记录有可能是别人的，也有可能是该用户历史上的。但它不能说出你为什么会喜欢的原因。难道大家都喜欢购买 A 和 B，就一定等于你买了 A 之后的果就是买 B 吗？未必，但的确需要承认，相关性很高——或者说，概率很大。

舍恩伯格认为，大数据时代只需要知道是什么，而无需知道为什么，就像亚马逊推荐算法一样，知道喜欢 A 的人很可能喜欢 B 但却不知道其中的原因。

1.2 大数据的产生和应用

人类历史上从未有哪个时代和今天一样产生如此海量的数据。数据的产生已经完全不受时间、地点的限制。从开始采用数据库作为数据管理的主要方式开始，人类社会的数据产生方式大致经历了 3 个阶段，而正是数据产生方式的巨大变化才最终导致大数据的产生。

1、运营式系统阶段。数据库的出现使得数据管理的复杂度大大降低，实际中数据库大都为运营系统所采用，作为运营系统的数据管理子系统。比如超市的销售记录系统，银行的交易记录系统、医院病人的医疗记录等。人类社会数据量第一次大的飞跃正是建立在运营式系统开始广泛使用数据库开始。这个阶段最主要特点是数据往往伴随着一定的运营活动而产生并记录在数据库中的，比如超市每销售出一件产品就会在数据库中产生相应的一条销售记录。这种数据的产生方式是被动的。

2、用户原创内容阶段。互联网的诞生促使人类社会数据量出现第二次大的飞跃。但是真正的数据爆发产生于 Web 2.0 时代，而 Web 2.0 的最重要标志就是用户原创内容（UGC,User Generated Content）。这类数据近几年一直呈现爆炸性的增长，主要有两个方面的原因。

首先是以博客、微博为代表的新型社交网络的出现和快速发展，使得用户产生数据的意愿更加强烈。其次就是以智能手机、平板电脑为代表的新型移动设备的出现，这些易携带、全天候接入网络的移动设备使得人们在网上发表自己意见的途径更为便捷。这个阶段数据的

产生方式是主动的。

3、感知式系统阶段。人类社会数据量第三次大的飞跃最终导致了大数据的产生，今天我们正处于这个阶段。这次飞跃的根本原因在于感知式系统的广泛使用。随着技术的发展，人们已经有能力制造极其微小的带有处理功能的传感器，并开始将这些设备广泛的布置于社会的各个角落，通过这些设备来对整个社会的运转进行监控。这些设备会源源不断的产生新数据，这种数据的产生方式是自动的。

简单来说，数据产生经历了被动、主动和自动三个阶段。这些被动、主动和自动的数据共同构成了大数据的数据来源，但其中自动式的数据才是大数据产生的最根本原因。

表 1-1 若干具有代表性的大数据应用及其特征

Applications	Examples	Number of Users	Response Time	Data Scale	Reliability	Accuracy
Scientific Computing	Bioinformatics	Small	Slow	TB	Moderate	Very High
Finance	High-frequency trading	Large	Very Fast	GB	Very High	Very High
Social network	Facebook	Very Large	Fast	PB	High	High
Mobile Data	Mobile phone	Very Large	Fast	TB	High	High
Internet of Things	Sensor network	Large	Fast	TB	High	High
Web Data	News website	Very Large	Fast	PB	High	High
Multimedia	Video site	Very Large	Fast	PB	High	Moderate

正如 Google 的首席经济学家 Hal Varian 所说，数据是广泛可用的，所缺乏的是从中提取出知识的能力。数据收集的根本目的是根据需求从数据中提取有用的知识，并将其应用到具体的领域之中。不同领域的大数据应用有不同的特点，表 1-1 列举了若干具有代表性的大数据应用及其特征。

正是由于大数据的广泛存在，才使得大数据问题的解决很具挑战性。而它的广泛应用，则促使越来越多的人开始关注和研究大数据问题。

1.3 大数据作用

大数据时代已经到来，认同这一判断的人越来越多。那么大数据意味着什么，他到底会改变什么？仅仅从技术角度回答，已不足以解惑。大数据只是宾语，离开了人这个主语，它再大也没有意义。我们需要把大数据放在人的背景中加以透视，理解它作为时代变革力量的所以然。

（1） 变革价值的力量

未来十年，决定中国是不是有大智慧的核心意义标准（那个"思想者"），就是国民幸福。一体现在民生上，二体现在生态上，通过大数据让有意义的事变得明晰，看我们在人与人关系上，做得是否比以前更有意义。总之，让我们从前 10 年的意义混沌时代，进入未来 10 年意义明晰时代。

（2） 变革经济的力量

生产者是有价值的，消费者是价值的意义所在。有意义的才有价值，消费者不认同的，就卖不出去，就实现不了价值；只有消费者认同的，才卖得出去，才实现得了价值。大数据帮助我们从消费者这个源头识别意义，从而帮助生产者实现价值。这就是启动内需的原理。

（3） 变革组织的力量

随着具有语义网特征的数据基础设施和数据资源发展起来，组织的变革就越来越显得不可避免。大数据将推动网络结构产生无组织的组织力量。最先反映这种结构特点的，是各种各样去中心化的 WEB2.0 应用，如 RSS、维基、博客等。 大数据之所以成为时代变革力量，在于它通过追随意义而获得智慧。

1.4 大数据与大规模数据、海量数据的差别

从对象角度看，大数据是大小超出典型数据库软件采集、储存、管理和分析等能力的数据集合。需要注意的是，大数据并非大量数据的简单无意义的堆积，数据量大并不意味着一定具有可观的利用前景。由于最终目标是从大数据中获取更多有价值的“新”信息，所以必然要求这些大量的数据之间存在着或远或近、或直接或间接的关联性，才具有相当的分析挖掘价值。数据间是否具有结构性和关联性，是 “大数据”与“大规模数据”的重要差别。

从技术角度看，大数据技术是从各种各样类型的大数据中，快速获得有价值信息的技术及其集成。“大数据”与“大规模数据”、“海量数据”等类似概念间的最大区别，就在于“大数据”这一概念中包含着对数据对象的处理行为。为了能够完成这一行为，从大数据对象中快速挖掘更多有价值的信息，使大数据“活起来”，就需要综合运用灵活的、多学科的方法，包括数据聚类、数据挖掘、分布式处理等，而这就需要拥有对各类技术、各类软硬件的集成应用能力。可见，大数据技术是使大数据中所蕴含的价值得以发掘和展现的重要工具。

从应用角度看，大数据是对特定的大数据集合、集成应用大数据技术、获得有价值信息的行为。正由于与具体应用紧密联系，甚至是一对一的联系，才使得“应用”成为大数据不可或缺的内涵之一。

需要明确的是，大数据分析处理的最终目标，是从复杂的数据集合中发现新的关联规则，继而进行深度挖掘，得到有效用的新信息。如果数据量不小，但数据结构简单，重复性高，分析处理需求也仅仅是根据已有规则进行数据分组归类，未与具体业务紧密结合，依靠已有基本数据分析处理技术已足够，则不能算作是完全的“大数据”，只是“大数据”的初级发展阶段。

1.5 典型的大数据应用实例

1.5.1 从谷歌流感趋势看大数据的应用价值

谷歌有一个名为“谷歌流感趋势”的工具，它通过跟踪搜索词相关数据来判断全美地区的流感情况（比如患者会搜索流感两个字）。近日，这个工具发出警告，全美的流感已经进入“紧张”级别。它对于健康服务产业和流行病专家来说是非常有用的，因为它的时效性极强，能够很好地帮助到疾病暴发的跟踪和处理。事实也证明，通过海量搜索词的跟踪获得的趋势报告是很有说服力的，仅波士顿地区，就有 700 例流感得到确认，该地区目前已宣布进入公共健康紧急状态。

这个工具工作的原理大致是这样的：设计人员置入了一些关键词（比如温度计、流感症状、肌肉疼痛、胸闷等），只要用户输入这些关键词，系统就会展开跟踪分析，创建地区流感图表和流感地图。谷歌多次把测试结果（蓝线）与美国疾病控制和预防中心的报告（黄线）做比对，从图 1-1 可知，两者结论存在很大相关性。

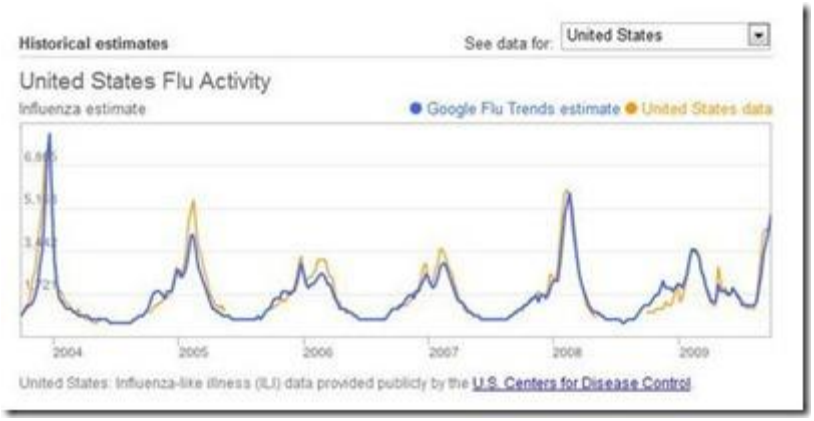

图 1-1 谷歌把测试结果（蓝线）与美国疾病控制和预防中心的报告（黄线）做比对

但它比线下收集的报告强在“时效性”上，因为患者只要一旦自觉有流感症状，在搜索和去医院就诊这两件事上，前者通常是他首先会去做的。就医很麻烦而且价格不菲，如果能自己通过搜索来寻找到一些自我救助的方案，人们就会第一时间使用搜索引擎。故而，还存

在一种可能是，医院或官方收集到的病例只能说明一小部分重病患者，轻度患者是不会去医院而成为它们的样本的。

1.5.2 大数据在医疗行业的应用

Seton Healthcare 是采用 IBM 最新沃森技术医疗保健内容分析预测的首个客户。该技术允许企业找到大量病人相关的临床医疗信息，通过大数据处理，更好地分析病人的信息。

在加拿大多伦多的一家医院，针对早产婴儿，每秒钟有超过 3000 次的数据读取。通过这些数据分析，医院能够提前知道哪些早产儿出现问题并且有针对性地采取措施，避免早产婴儿夭折。

大数据让更多的创业者更方便地开发产品，比如通过社交网络来收集数据的健康类 App。也许未来数年后，它们搜集的数据能让医生给你的诊断变得更为精确，比方说不是通用的成人每日三次一次一片，而是检测到你的血液中药剂已经代谢完成会自动提醒你再次服药。

1.5.3 大数据在能源行业的应用

智能电网现在欧洲已经做到了终端，也就是所谓的智能电表。在德国，为了鼓励利用太阳能，会在家庭安装太阳能，除了卖电给你，当你的太阳能有多余电的时候还可以买回来。通过电网收集每隔五分钟或十分钟收集一次数据，收集来的这些数据可以用来预测客户的用电习惯等，从而推断出在未来 2~3 个月时间里，整个电网大概需要多少电。有了这个预测后，就可以向发电或者供电企业购买一定数量的电。因为电有点像期货一样，如果提前买就会比较便宜，买现货就比较贵。通过这个预测后，可以降低采购成本。

维斯塔斯风力系统，依靠的是 BigInsights 软件和 IBM 超级计算机，然后对气象数据进行分析，找出安装风力涡轮机和整个风电场最佳的地点。利用大数据，以往需要数周的分析工作，现在仅需要不足 1 小时便可完成。

1.5.4 大数据在通信行业的应用

XO Communications 通过使用 IBM SPSS 预测分析软件，减少了将近一半的客户流失率。XO 现在可以预测客户的行为，发现行为趋势，并找出存在缺陷的环节，从而帮助公司及时采取措施，保留客户。此外，IBM 新的 Netezza 网络分析加速器，将通过提供单个端到端网

络、服务、客户分析视图的可扩展平台，帮助通信企业制定更科学、合理决策。

电信业者透过数以千万计的客户资料，能分析出多种使用者行为和趋势，卖给需要的企业，这是全新的资料经济。

中国移动通过大数据分析，对企业运营的全业务进行针对性的监控、预警、跟踪。系统在第一时间自动捕捉市场变化，再以最快捷的方式推送给指定负责人，使他在最短时间内获知市场行情。

NTT docomo 把手机位置信息和互联网上的信息结合起来，为顾客提供附近的餐饮店信息，接近末班车时间时，提供末班车信息服务。

1.5.5 大数据在零售业的应用

"我们的某个客户，是一家领先的专业时装零售商，通过当地的百货商店、网络及其邮购目录业务为客户提供服务。公司希望向客户提供差异化服务，如何定位公司的差异化，他们通过从 Twitter 和 Facebook 上收集社交信息，更深入的理解化妆品的营销模式，随后他们认识到必须保留两类有价值的客户：高消费者和高影响者。希望通过接受免费化妆服务，让用户进行口碑宣传，这是交易数据与交互数据的完美结合，为业务挑战提供了解决方案。"Informatica 的技术帮助这家零售商用社交平台上的数据充实了客户主数据，使他的业务服务更具有目标性。

零售企业也监控客户的店内走动情况以及与商品的互动。它们将这些数据与交易记录相结合来展开分析，从而在销售哪些商品、如何摆放货品以及何时调整售价上给出意见，此类方法已经帮助某领先零售企业减少了 17%的存货，同时在保持市场份额的前提下，增加了高利润率自有品牌商品的比例。

1.6 从数据库到大数据

大数据的出现，必将颠覆传统的数据管理方式。在数据来源、数据处理方式和数据思维等方面都会对其带来革命性的变化。本书作者主要从事数据库领域的研究，因此，编写本书时，主要侧重于从数据库存储和管理方面介绍大数据技术。对于数据库研究人员和从业人员而言，必须清楚的是，从数据库(DB)到大数据(BD)，看似只是一个简单的技术演进，但细细考究不难发现两者有着本质上的差别。

如果要用简单的方式来比较传统的数据库和大数据的区别的话，我们认为“池塘捕鱼”

和“大海捕鱼”是个很好的类比。“池塘捕鱼”代表着传统数据库时代的数据管理方式，而“大海捕鱼”则对应着大数据时代的数据管理方式，“鱼”是待处理的数据。“捕鱼”环境条件的变化导致了“捕鱼”方式的根本性差异。这些差异主要体现在如下几个方面：

1、**数据规模**：“池塘”和“大海”最容易发现的区别就是规模。“池塘”规模相对较小，即便是先前认为比较大的“池塘”，譬如 VLDB(Very Large Database)，和“大海”XLDB(Extremely Large Database)相比仍旧偏小。“池塘”的处理对象通常以 MB 为基本单位，而“大海”则常常以 GB，甚至是 TB、PB 为基本处理单位。

2、**数据类型**：过去的“池塘”中，数据的种类单一，往往仅仅有一种或少数几种，这些数据又以结构化数据为主。而在“大海”中，数据的种类繁多，数以千计，而这些数据又包含着结构化、半结构化以及非结构化的数据，并且半结构化和非结构化数据所占份额越来越大。

3、**模式(Schema)和数据的关系**：传统的数据库都是先有模式，然后才会产生数据。这就好比是先选好合适的“池塘”，然后才会向其中投放适合在该“池塘”环境生长的“鱼”。而大数据时代很多情况下难以预先确定模式，模式只有在数据出现之后才能确定，且模式随着数据量的增长处于不断的演变之中。这就好比先有少量的鱼类，随着时间推移，鱼的种类和数量都在不断的增长。鱼的变化会使大海的成分和环境处于不断的变化之中。

4、**处理对象**：在“池塘”中捕鱼，“鱼”仅仅是其捕捞对象。而在“大海”中，“鱼”除了是捕捞对象之外，还可以通过某些“鱼”的存在来判断其他种类的“鱼”是否存在。也就是说传统数据库中数据仅作为处理对象。而在大数据时代，要将数据作为一种资源来辅助解决其他诸多领域的问题。

5、**处理工具**：捕捞“池塘”中的“鱼”，一种渔网或少数几种基本就可以应对，也就是所谓的 One Size Fits All。但是在“大海”中，不可能存在一种渔网能够捕获所有的鱼类，也就是说 No Size Fits All。

从“池塘”到“大海”，不仅仅是规模的变大。传统的数据库代表着数据工程(Data Engineering)的处理方式，大数据时代的数据已不仅仅只是工程处理的对象，需要采取新的数据思维来应对。图灵奖获得者、著名数据库专家 Jim Gray 博士观察并总结人类自古以来，在科学研究上，先后历经了实验、理论和计算三种范式。当数据量不断增长和累积到今天，传统的三种范式在科学研究，特别是一些新的研究领域已经无法很好的发挥作用，需要有一种全新的第四种范式来指导新形势下的科学研究。基于这种考虑，Jim Gray 提出了一种新的数据探索型研究方式，被他自己称之为科学研究的“第四种范式”(The Fourth Paradigm)。

表 1-2 四种范式的比较

Science Paradigms	Time	Methodology
Empirical	Thousand years ago	Describing natural phenomena
Theoretical	Last few hundred years	Using models, generalizations
Computational	Last few decades	Simulating complex phenomena
Data Exploration (eScience)	Today	Data captured by instruments or generated by simulator; Processed by software; Information stored in computer; Scientist analyzes database

四种范式的比较如表 1-2 所示。第四种范式的实质就是从以计算为中心，转变到以数据处理为中心，也就是我们所说的数据思维。这种方式需要我们从根本上转变思维。正如前面提到的“捕鱼”，在大数据时代，数据不再仅仅是“捕捞”的对象，而应当转变成一种基础资源，用数据这种资源来协同解决其他诸多领域的问题。计算社会科学(Computational Social Science)基于特定社会需求，在特定的社会理论指导下，收集、整理和分析数据足迹(data print)，以便进行社会解释、监控、预测与规划的过程和活动。计算社会科学是一种典型的需要采用第四种范式来做指导的科学研究领域。Duncan J. Watts 在《自然》杂志上的文章《A twenty-first century science》也指出借助于社交网络和计算机分析技术，21 世纪的社会科学有可能实现定量化的研究，从而成为一门真正的自然科学。

1.7 大数据与云计算

近几年来，云计算受到学术界和工业界的热捧，随后，大数据横空出世，更是炙手可热。那么，大数据和云计算之间是什么关系呢？

（1）从整体上看，大数据与云计算是相辅相成的

大数据着眼于“数据”，关注实际业务，提供数据采集分析挖掘，看重的是信息积淀，即数据存储能力。云计算着眼于“计算”，关注 IT 解决方案，提供 IT 基础架构，看重的是计算能力，即数据处理能力。没有大数据的信息积淀，则云计算的计算能力再强大，也难以找到用武之地；没有云计算的处理能力，则大数据的信息积淀再丰富，也终究只是镜花水月。

（2）从技术上看，大数据根植于云计算

云计算关键技术中的海量数据存储技术、海量数据管理技术、MapReduce 编程模型，都是大数据技术的基础。

表 1-3 云计算和大数据技术的关系

	云计算技术	描述
大数据的关键技术	虚拟化技术	软硬件隔离，资源整合
	云计算平台管理技术	大规模系统运营，快速故障检测与恢复
	MapReduce编程模型	分布式编程模型，用于并行处理大规模数据集的软件框架
	海量数据存储技术	分布式存储方式存储数据，冗余存储方式保证系统可靠
	海量数据管理技术	NoSQL数据库，进行海量数据管理以便后续分析挖掘

（3）大数据技术与云计算有相同，也有差异

表 1-4 云计算和大数据技术的异同

		大数据	云计算
总体关系		云计算为大数据提供了有力的工具和途径，大数据为云计算提供了很有价值的用武之地	
相同点		1. 都是为数据存储和处理服务 2. 都需要占用大量的存储和计算资源，因而都要用到海量数据存储技术、海量数据管理技术、MapReduce等并行处理技术	
差异点	背景	现有的数据处理技术不能胜任社交网络和物联网产生的大量异构数据，但这些数据存在很大价值	基于互联网的相关服务日益丰富和频繁
	目的	充分挖掘海量数据中的信息	通过互联网更好地调用、扩展和管理计算及存储方面的资源和能力
	对象	数据	IT资源、能力和应用
	推动力量	从事数据存储与处理的软件厂商和拥有大量数据的企业	生产计算及存储设备的厂商、拥有计算及存储资源的企业
	带来的价值	发现数据中的价值	节省IT部署成本

（4）大数据技术与云计算相结合会带来什么？

图 1-2 大数据与云计算的结合

（5）大数据的商业模式与架构----云计算及其分布式结构是重要途径

大数据处理技术正在改变目前计算机的运行模式，正在改变着这个世界。它能处理几乎各种类型的海量数据，无论是微博、文章、电子邮件、文档、音频、视频，还是其它形态的数据。它工作的速度非常快速，实际上几乎实时。它具有普及性，因为它所用的都是最普通低成本的硬件，而云计算将计算任务分布在大量计算机构成的资源池上，使用户能够按需获取计算力、存储空间和信息服务。云计算及其技术给了人们廉价获取巨量计算和存储的能力，云计算分布式架构能够很好地支持大数据存储和处理需求。这样的低成本硬件+低成本软件+低成本运维，更加经济和实用，使得大数据处理和利用成为可能。

1.8 大数据与物联网

物联网是新一代信息技术的重要组成部分，其英文名称是“The Internet of things”。顾名思义，“物联网就是物物相连的互联网”。这有两层意思：第一，物联网的核心和基础仍然是互联网，是在互联网基础上的延伸和扩展的网络；第二，其用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信。物联网就是“物物相连的互联网”。物联网通过智能感知、识别技术与普适计算、泛在网络的融合应用，被称为继计算机、互联网之后世界信息产业发展的第三次浪潮。物联网是互联网的应用拓展，与其说物联网是网络，不如说物联网

是业务和应用。因此，应用创新是物联网发展的核心，以用户体验为核心的创新 2.0 是物联网发展的灵魂。

物联网架构可分为三层，包括感知层、网络层和应用层：

- 感知层：由各种传感器构成，包括温湿度传感器、二维码标签、RFID 标签和读写器、摄像头、GPS 等感知终端。感知层是物联网识别物体、采集信息的来源；
- 网络层：由各种网络，包括互联网、广电网、网络管理系统和云计算平台等组成，是整个物联网的中枢，负责传递和处理感知层获取的信息；
- 应用层：是物联网和用户的接口，它与行业需求结合，实现物联网的智能应用。

物联网用途广泛，遍及智能交通、环境保护、政府工作、公共安全、平安家居、智能消防、工业监测、环境监测、路灯照明管控、景观照明管控、楼宇照明管控、广场照明管控、老人护理、个人健康、花卉栽培、水系监测、食品溯源、敌情侦查和情报搜集等多个领域。

国际电信联盟于 2005 年的报告曾描绘“物联网”时代的图景：当司机出现操作失误时汽车会自动报警；公文包会提醒主人忘带了什么东西；衣服会“告诉”洗衣机对颜色和水温的要求等等。物联网在物流领域内的应用则比如：一家物流公司应用了物联网系统的货车，当装载超重时，汽车会自动告诉你超载了，并且超载多少，但空间还有剩余，告诉你轻重货怎样搭配；当搬运人员卸货时，一只货物包装可能会大叫“你扔疼我了”，或者说“亲爱的，请你不要太野蛮，可以吗？”；当司机在和别人扯闲话，货车会装作老板的声音怒吼“笨蛋，该发车了！”

物联网把新一代 IT 技术充分运用在各行各业之中，具体地说，就是把感应器嵌入和装备到电网、铁路、桥梁、隧道、公路、建筑、供水系统、大坝、油气管道等各种物体中，然后将“物联网”与现有的互联网整合起来，实现人类社会与物理系统的整合，在这个整合的网络当中，存在能力超级强大的中心计算机集群，能够对整合网络内的人员、机器、设备和基础设施实施实时的管理和控制，在此基础上，人类可以以更加精细和动态的方式管理生产和生活，达到“智慧”状态，提高资源利用率和生产力水平，改善人与自然间的关系。

物联网，移动互联网再加上传统互联网，每天都在产生海量数据，而大数据又通过云计算的形式，将这些数据筛选处理分析，提取出有用的信息，这就是大数据分析。

1.9 对大数据的错误认识

大数据对于悲观者而言，意味着数据存储世界的末日，对乐观者而言，这里孕育了巨

大的市场机会，庞大的数据就是一个信息金矿，随着技术的进步，其财富价值将很快被我们发现，而且越来越容易。

随着物联网和云计算的研究和应用不断深入，对大数据的研究越来越引起广泛的重视，对大数据及其处理技术产生了很多错误的认识，业界有大量关于何谓大数据及它可以做什么的说法，其中有很多是相互矛盾的，都存在一定的片面性，根据 IDC2011 年市场研究报告，主要有三个典型的错误说法：

1） 关系型数据库不能扩展到非常大的数据量，因此不被认为是大数据的技术；

2） 无论工作负载有多大，也无论使用场景如何，Hadoop（或推而广之，任何 Mapreduce 的环境）都是大数据的最佳选择；

3） 基于数据模型的数据库管理系统的时代已经结束了，数据模型必须大数据的方式来建立。

正确的结论是，新型关系型数据库既可解决结构化和非结构化数据，也可满足大数据的数量和速度要求，相比较而言，Hadoop 型解决方案是片面的，不能解决很多的关系型应用环境问题，不一定是最佳选择，大数据管理和处理有更优的解决方案和技术路线。

1.10 大数据技术

大数据本身是一个现象而不是一种技术，伴随着大数据的采集、传输、处理和应用的相关技术就是大数据处理技术，是一系列使用非传统的工具来对大量的结构化、半结构化和非结构化数据进行处理，从而获得分析和预测结果的一系列数据处理技术，或简称大数据技术。

大数据技术主要包括：

- 数据采集：ETL 工具负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。
- 数据存取：关系数据库、NoSQL、SQL 等。
- 基础架构：云存储、分布式文件存储等。
- 数据处理：自然语言处理(NLP，Natural Language Processing)是研究人与计算机交互的语言问题的一门学科。处理自然语言的关键是要让计算机"理解"自然语言，所以自然语言处理又叫做自然语言理解(NLU，NaturalLanguage Understanding)，也称为

计算语言学(Computational Linguistics。一方面它是语言信息处理的一个分支，另一方面它是人工智能(AI， Artificial Intelligence)的核心课题之一。

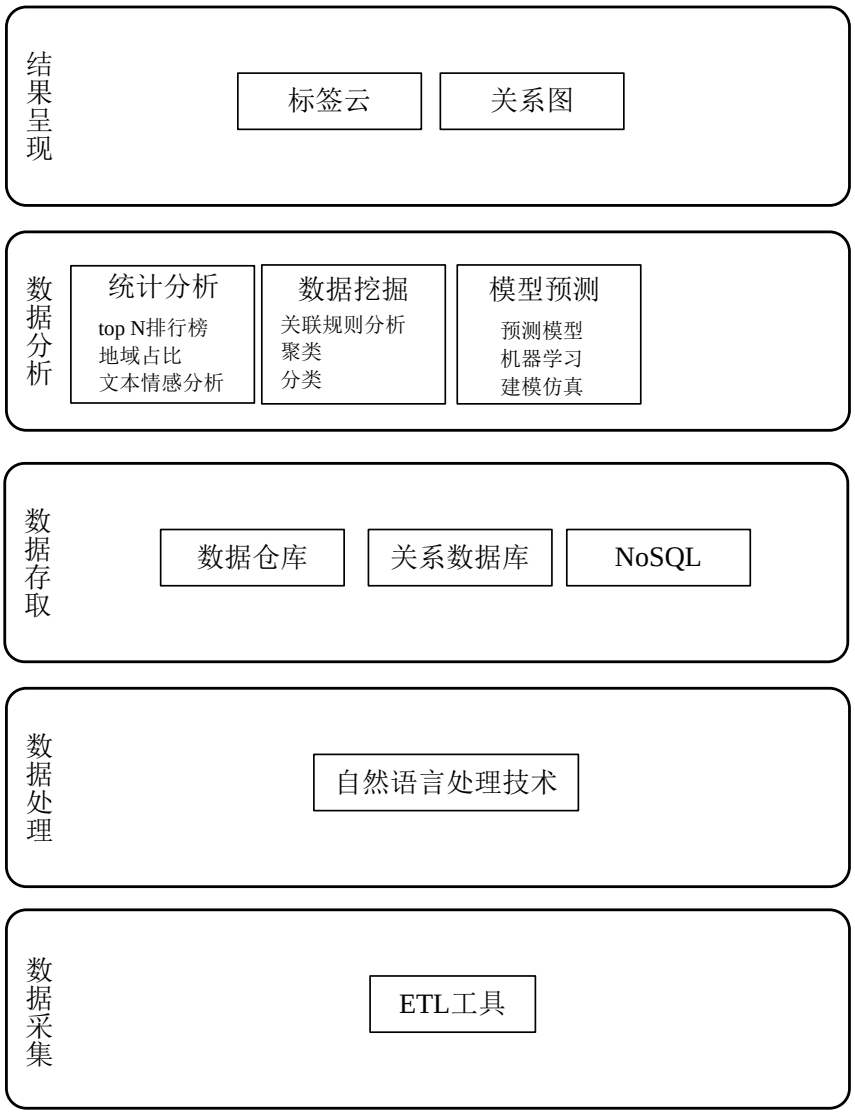

图 1-3 大数据技术内容框架图

- 统计分析：假设检验、显著性检验、差异分析、相关分析、T 检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、岭回归、logistic 回归分析、曲线估计、因子分析、聚类分析、主成分分析、因子分析、快速聚类法与聚类法、判别分析、对应分析、多元对应分析（最优尺度分析）、bootstrap 技术等等。
- 数据挖掘：分类 （Classification）、估计（Estimation）、预测（Prediction）、相关性分组或关联规则（Affinity grouping or association rules）、聚类（Clustering）、描述和

可视化、Description and Visualization）、复杂数据类型挖掘(Text， Web ，图形图像，视频，音频等)。

- 模型预测：预测模型、机器学习、建模仿真。
- 结果呈现：云计算、标签云、关系图等。

1.11 大数据存储和管理技术

Big Data（大数据技术）是近来的一个技术热点，但从名字就能判断它并不是什么新词。毕竟，大是一个相对概念。历史上，数据库、数据仓库、数据集市等信息管理领域的技术，很大程度上也是为了解决大规模数据的问题。被誉为数据仓库之父的 Bill Inmon 早在 20 世纪 90 年代就经常将 Big Data 挂在嘴边了。

然而，Big Data 作为一个专有名词成为热点，主要应归功于近年来互联网、云计算、移动和物联网的迅猛发展。无所不在的移动设备、RFID、无线传感器每分每秒都在产生数据，数以亿计用户的互联网服务时时刻刻在产生巨量的交互……要处理的数据量实在是太大、增长太快了，而业务需求和竞争压力对数据处理的实时性、有效性又提出了更高要求，传统的常规技术手段根本无法应付。

在这种情况下，技术人员纷纷研发和采用了一批新技术，主要包括分布式缓存、基于 MPP 的分布式数据库、分布式文件系统、各种 NoSQL 分布式存储方案等。

1.11.1分布式缓存

分布式缓存使用 CARP（Caching Array Routing Protocol）技术，可以产生一种高效率无接缝式的缓存，使用上让多台缓存服务器形同一台，并且不会造成数据重复存放的情况。分布式缓存提供的数据内存缓存可以分布于大量单独的物理机器中。换句话说，分布式缓存所管理的机器实际上就是一个集群。它负责维护集群中成员列表的更新，并负责执行各种操作，比如说在集群成员发生故障时执行故障转移，以及在机器重新加入集群时执行故障恢复。

分布式缓存支持一些基本配置：重复(replicated)、分区(partitioned)和分层(tiered)。重复(Replication)用于提高缓存数据的可用性。在这种情况下，数据将重复缓存在分布式系统的多台成员机器上，这样只要有一个成员发生故障，其他成员便可以继续处理该数据的提供。另一方面，分区(Partitioning)是一种用于实现高可伸缩性的技巧。通过将数据分区存放在许

多机器上，内存缓存的大小加随着机器的增加而呈线性增长。结合分区和重复这两种机制创建出的缓存可同时具备大容量和高可伸缩的特性。分层缓存也称作客户机-服务器(client-server)缓存，它是一种拓扑结构，在该结构中缓存功能将集中于一组机器上。缓存客户机通常并不会亲自执行任何缓存操作，而是连接到缓存并检索或更新其中的数据。分层缓存架构可以包含多层结构。

mem-cached 是 danga.com（运营 LiveJournal 的技术团队）开发的一套分布式内存对象缓存系统，用于在动态系统中减少数据库负载，提升性能。许多 Web 应用程序都将数据保存到 RDBMS 中，应用服务器从中读取数据并在浏览器中显示。但随着数据量的增大，访问的集中，就会出现 REBMS 的负担加重，数据库响应恶化，网站显示延迟等重大影响。Memcached 是高性能的分布式内存缓存服务器。一般的使用目的是通过缓存数据库查询结果，减少数据库的访问次数，以提高动态 Web 应用的速度、提高扩展性。

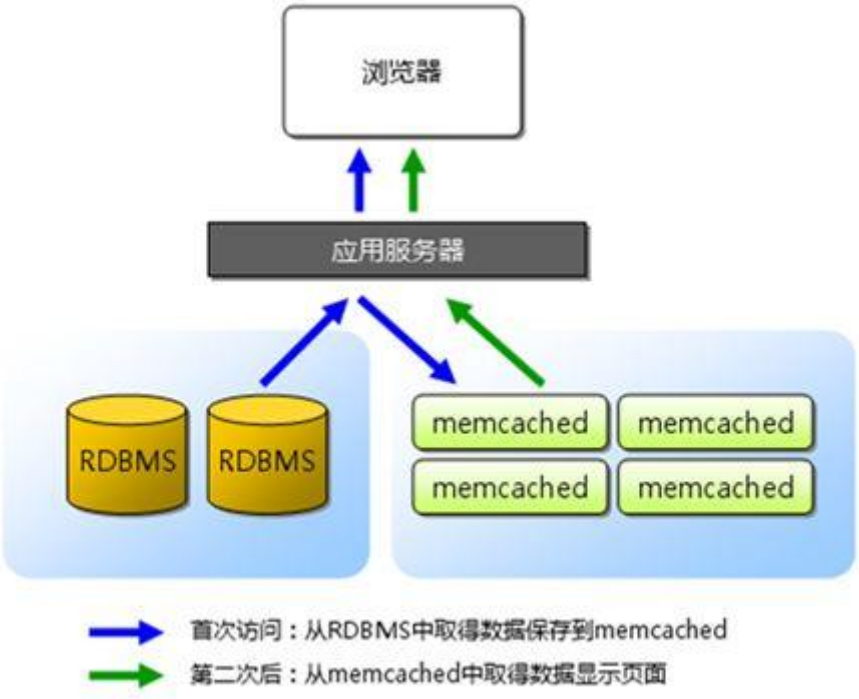

图 1-4 memcached 示意图

Memcached 作为高速运行的分布式缓存服务器具有以下特点：

- 协议简单：memcached 的服务器客户端通信并不使用复杂的 MXL 等格式，而是使用简单的基于文本的协议。
- 基于 libevent 的事件处理：libevent 是个程序库，他将 Linux 的 epoll、BSD 类操作系统的 kqueue 等时间处理功能封装成统一的接口。memcached 使用这个 libevent 库，因此能在 Linux、BSD、Solaris 等操作系统上发挥其高性能。

- 内置内存存储方式：为了提高性能，memcached 中保存的数据都存储在 memcached 内置的内存存储空间中。由于数据仅存在于内存中，因此重启 memcached，重启操作系统会导致全部数据消失。另外，内容容量达到指定的值之后 memcached 回自动删除不适用的缓存。
- Memcached 不互通信的分布式：memcached 尽管是“分布式”缓存服务器，但服务器端并没有分布式功能。各个 memcached 不会互相通信以共享信息。它的分布式主要是通过客户端实现的。

memcached 处理的原子是每一个（Key，Value）对（以下简称 KV 对），Key 会通过一个 hash 算法转化成 hash-Key，便于查找、对比以及做到尽可能的散列。同时，memcached 用的是一个二级散列，通过一张大 hash 表来维护。

memcached 由两个核心组件组成：服务端（ms）和客户端（mc），在一个 memcached 的查询中，ms 先通过计算 Key 的 hash 值来确定 KV 对所处在的 ms 位置。当 ms 确定后，mc 就会发送一个查询请求给对应的 ms，让它来查找确切的数据。因为这之间没有交互以及多播协议，所以 memcached 交互带给网络的影响是最小化的。

MemcacheDB 是一个分布式、Key-Value 形式的持久存储系统。它不是一个缓存组件，而是一个基于对象存取的、可靠的、快速的持久存储引擎。协议与 memcached 一致（不完整），所以，很多 memcached 客户端都可以跟它连接。MemcacheDB 采用 Berkeley DB 作为持久存储组件，因此，很多 Berkeley DB 的特性它都支持。

类似这样的产品也很多，如淘宝 Tair 就是 Key-Value 结构存储，在淘宝得到了广泛使用。后来 Tair 也做了一个持久化版本，思路基本与新浪 MemcacheDB 一致。

1.11.2分布式数据库

分布式数据库系统通常使用较小的计算机系统，每台计算机可单独放在一个地方，每台计算机中都有 DBMS 的一份完整拷贝副本，并具有自己局部的数据库，位于不同地点的许多计算机通过网络互相连接，共同组成一个完整的、全局的大型数据库。

分布式数据库系统是数据库系统与计算机网络相结合的产物。分布式数据库系统产生于 1970 年代末期，在 1980 年代进入迅速成长阶段。由于数据库应用需求的拓展和计算机硬件环境的改变，计算机网络与数字通信技术的飞速发展，卫星通信、蜂窝通信、计算机局域网、广域网和 Internet 的迅速发展，使得分布式数据库系统应运而生，并成为计算机技术最活跃

的研究领域之一。

分布式数据库系统符合信息系统应用的需求，符合当前企业组织的管理思想和管理方式。对于地域上分散而管理上又相对集中的大企业而言，数据通常是分布存储在不同地理位置，每个部门都会负责维护与自己工作相关的数据。整个企业的信息就被分隔成多个“信息孤岛”。分布式数据库为这些信息孤岛提供了一座桥梁。分布式数据库的结构能够反映当今企业组织的信息数据结构，本地数据保存在本地维护，而又可以在需要时存取异地数据。也就是说，既需要有各部门的局部控制和分散管理，同时也需要整个组织的全局控制和高层次的协同管理。这种协同管理要求各部门之间的信息既能灵活交流与共享，又能统一管理和使用，自然而然就提出了对分布式数据库系统的需求。随着应用需求的扩大和要求的提高，人们越来越认识到集中式数据库的局限性，迫切需要把这些子部门的信息通过网络连接起来，组成一个分布式数据库。

世界上第一个分布式数据库系统 SDD-1，是由美国计算机公司于 1976 年-1978 年设计的，并于 1979 年在 DEC-10 和 DEC-20 计算机上面实现。

Spanner 是一个可扩展、多版本、全球分布式并支持同步复制的分布式数据库。它是 Google 的第一个可以全球扩展并且支持外部一致性事务的分布式数据库。Spanner 能做到这些，离不开一个用 GPS 和原子钟实现的时间 API。这个 API 能将数据中心之间的时间同步精确到 10ms 以内。因此，Spanner 有几个给力的功能：无锁读事务、 原子模式修改、读历史数据无阻塞。

1.11.3分布式文件系统

谈到分布式文件系统，不得不提的是 Google 的 GFS。基于大量安装有 Linux 操作系统的普通 PC 构成的集群系统，整个集群系统由一台 Master（通常有几台备份）和若干台 TrunkServer 构成。GFS 中文件被分成固定大小的 Trunk 分别存储在不同的 TrunkServer 上，每个 Trunk 有多份（通常为 3 份）拷贝，也存储在不同的 TrunkServer 上。Master 负责维护 GFS 中的 Metadata，即文件名及其 Trunk 信息。客户端先从 Master 上得到文件的 Metadata，根据要读取的数据在文件中的位置与相应的 TrunkServer 通信，获取文件数据。

在 Google 的论文发表后，就诞生了 Hadoop。截至今日，Hadoop 被很多中国最大互联网公司所追捧，百度的搜索日志分析，腾讯、淘宝和支付宝的数据仓库都可以看到 Hadoop 的身影。

Hadoop 具备低廉的硬件成本、开源的软件体系、较强的灵活性、允许用户自己修改代码等特点，同时能支持海量数据存储和计算任务。

1.11.4NoSQL

NoSQL 数据库，指的是非关系型的数据库。随着互联网 web2.0 网站的兴起，传统的关系数据库在应付 web2.0 网站，特别是超大规模和高并发的 SNS 类型的 web2.0 纯动态网站已经显得力不从心，暴露了很多难以克服的问题，而非关系型的数据库则由于其本身的特点得到了非常迅速的发展。

现今的计算机体系结构在数据存储方面要求具备庞大的水平扩展性（horizontal scalability，是指能够连接多个软硬件的特性，这样可以将多个服务器从逻辑上看成一个实体），而 NoSQL 致力于改变这一现状。目前 Google 的 BigTable 和 Amazon 的 Dynamo 使用的就是 NoSQL 型数据库。2010 年下半年，Facebook 选择 HBase 来做实时消息存储系统，替换原来开发的 Cassandra 系统。这使得很多人开始关注 NoSQL 型数据库 HBase。Facebook 选择 HBase 是基于短期小批量临时数据和长期增长的很少被访问到的数据这两个需求来考虑的。

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，利用 HBase 技术可在廉价 PC Server 上搭建大规模结构化存储集群。HBase 是 BigTable 的开源实现，使用 HDFS 作为其文件存储系统。Google 运行 MapReduce 来处理 BigTable 中的海量数据，HBase 同样利用 MapReduce 来处理 HBase 中的海量数据；BigTable 利用 Chubby 作为协同服务，HBase 则利用 Zookeeper 作为对应。

1.12 大数据生态系统

以下是福布斯专栏作家 Dave Feinleib 绘制的一张大数据生态系统图谱，非常有参考价值。该信息图中涉及的公司、产品和技术：

- Splunk,Loggly,Sumologic
- Predictive Policing,BloomReach
- Gnip,Datasift,Space Curve,Inrix
- Oracle Hyperion, SAP BusinessObjects，Microsoft Business Intelligence, IBM Cognos, SAS,MicroStrategy, GoodData
- Tableau Software, Palantir, MetaMarkets, Teradata Aster, Visual.ly, KarmaSphere, EMC

Greenplum, Platfora, ClearStory

- HortonWorks, Cloudera, MapR, Vertica
- Couchbase, Teradata, 10gen, Hadapt
- Amazon Web Services Elastic MapReduce, Infochimps, Microsoft Windows Azure
- Oracle, Microsoft SQL Server, MySQL, PostgreSQL, memsql
- Hadoop, MapReduce, Hbase, Cassandra

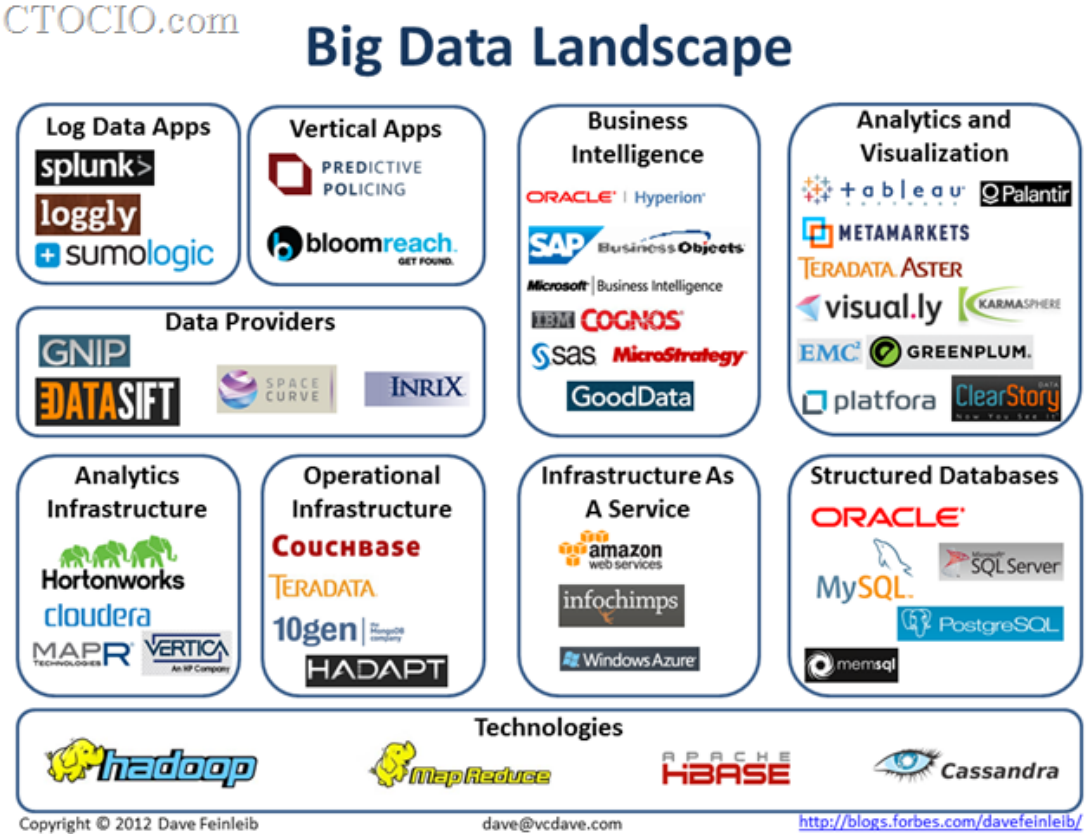

图 1-5 大数据生态系统

本章小结

本章介绍了大数据的概念、产生、应用与作用，并介绍了大数据与大规模数据、海量数据的差别，同时给出了大型的大数据应用案例；同时，分析了大数据与云计算、物联网的相互关系，并提到了一些关于大数据的错误认识；介绍了大数据技术以及大数据存储与管理技术；最后描述了大数据生态系统。

参考文献

[1] 孟小峰，慈祥. 大数据管理：概念、技术与挑战. 计算机学报, 2013 年第 8 期.

[2] 关志刚. 信息图：大数据企业生态图谱. IT 经理网.

http://www.ctocio.com/bigdata/7028.html

[3] 百度百科. 物联网.

[4] 邵佩英. 分布式数据库系统及其应用，科学出版社.

[5] 其他网络来源.

附录 1:任课教师介绍

林子雨(1978－),男,博士,厦门大学计算机科学系助理教授,主要研究领域为数据库,数据仓库,数据挖掘.
主讲课程：《大数据技术基础》
办公地点：厦门大学海韵园科研 2 号楼
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu