厦门大学计算机科学系研究生课程

《大数据技术基础》

第1章 大数据概述

（2013年新版）

林子雨

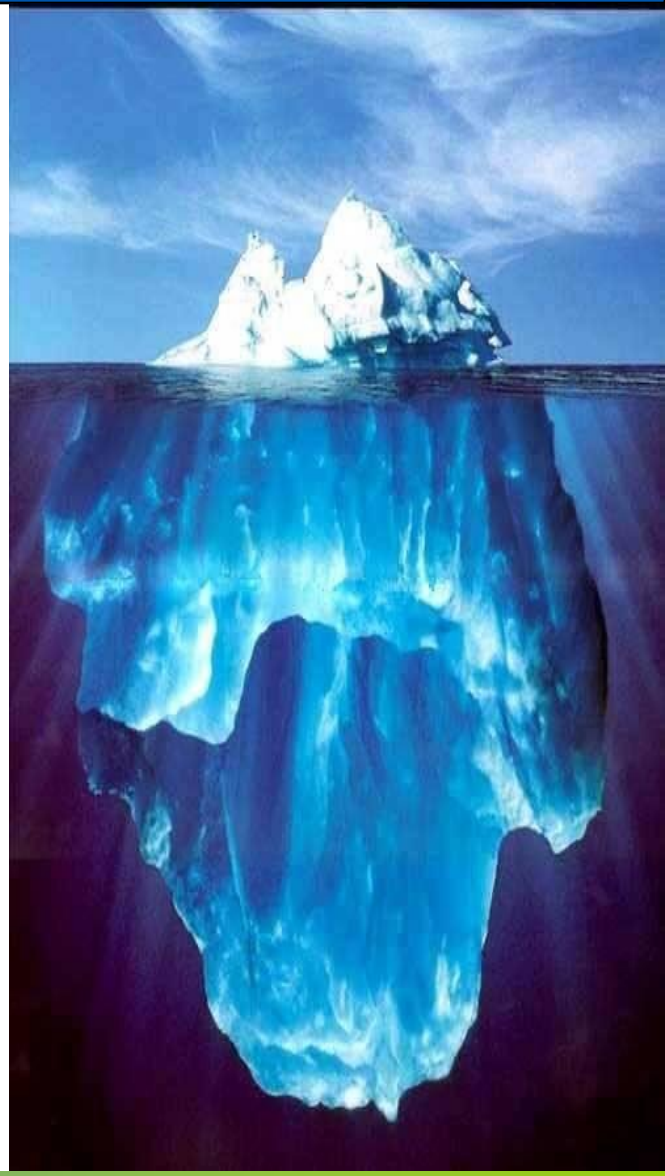
厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

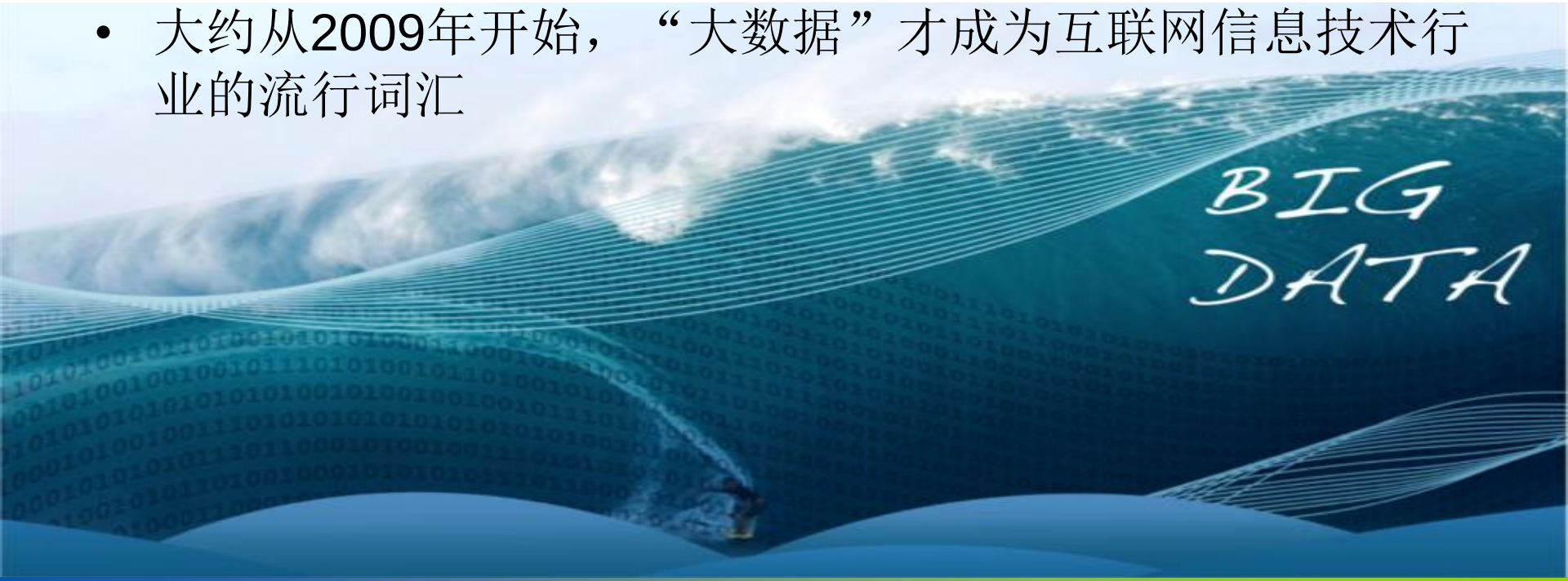
提纲

- 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、海量数据的区别
- 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



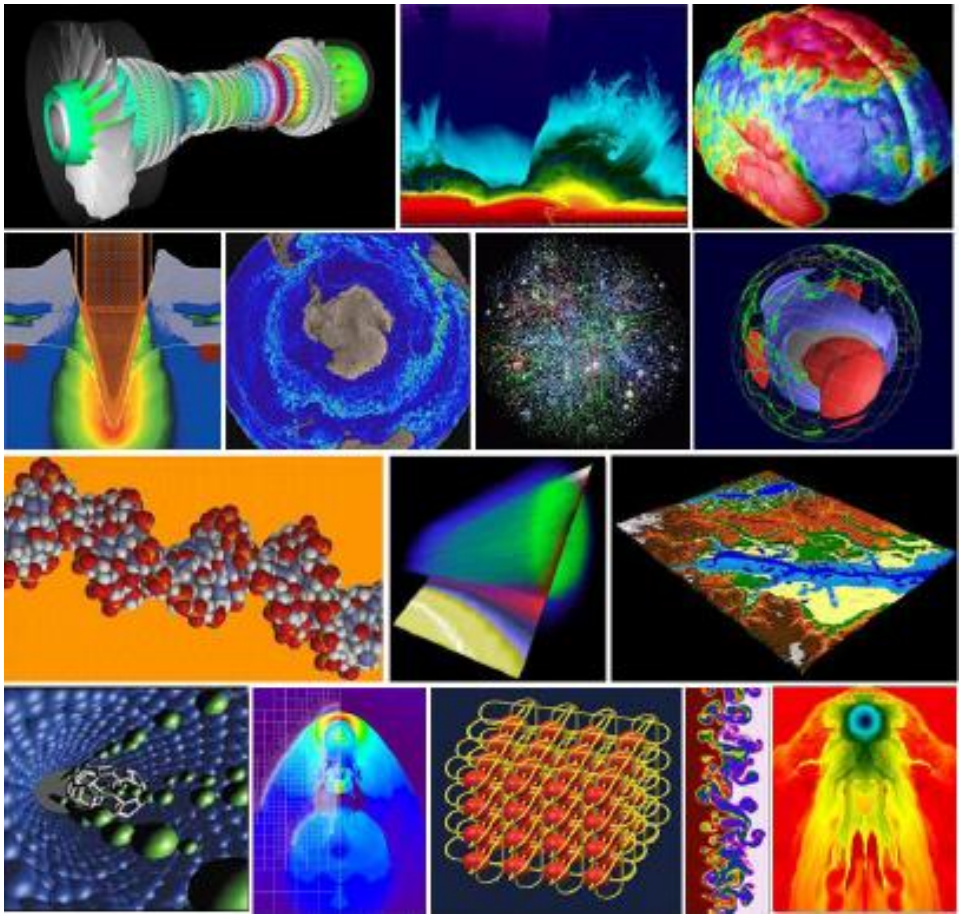
大数据

- “大数据”是时下最火热的IT行业词汇
- 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”。
- 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

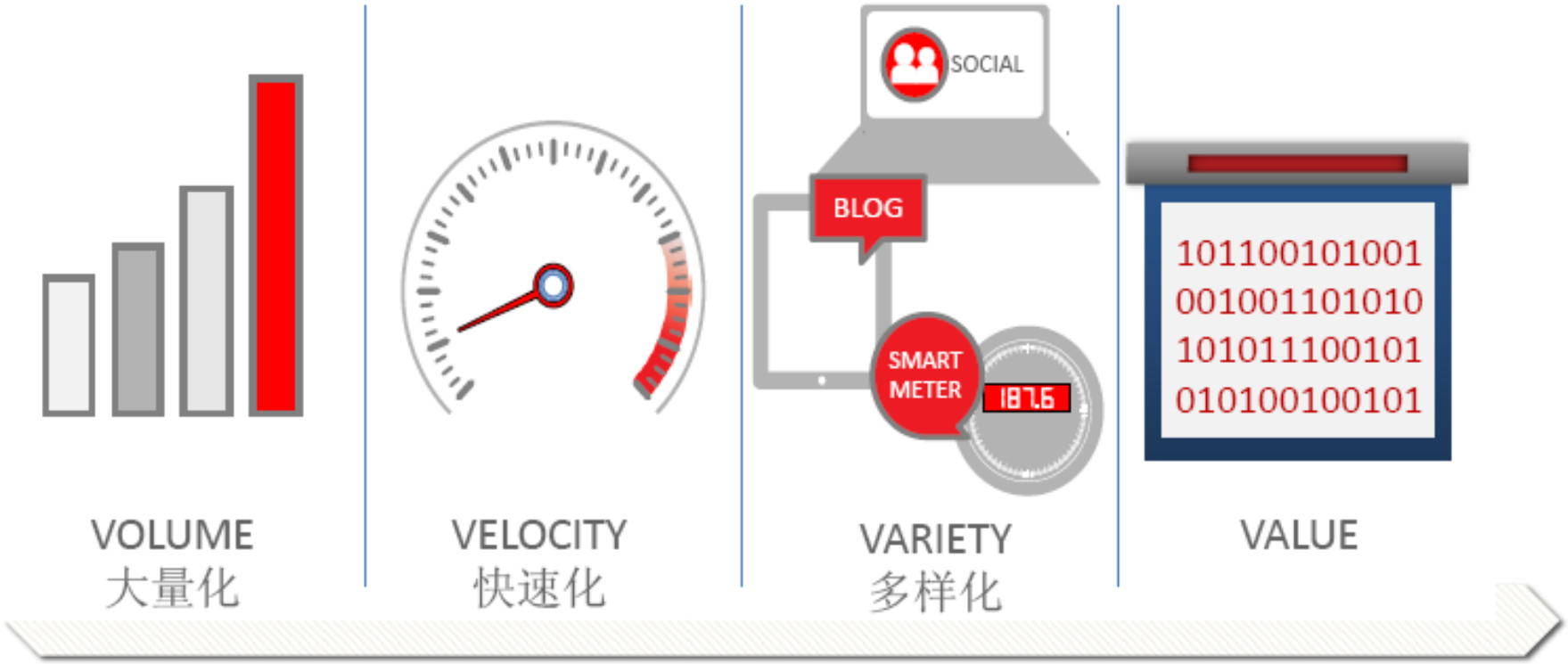


大数据无处不在

- 科学研究
  - 基因组
  - LHC 加速器
  - 地球与空间探测
- 企业应用
  - Email、文档、文件
  - 应用日志
  - 交易记录
- Web 1.0数据
  - 文本
  - 图像
  - 视频
- **Web 2.0**数据
  - 查询日志/点击流
  - **Twitter/ Blog / SNS**
  - **Wiki**



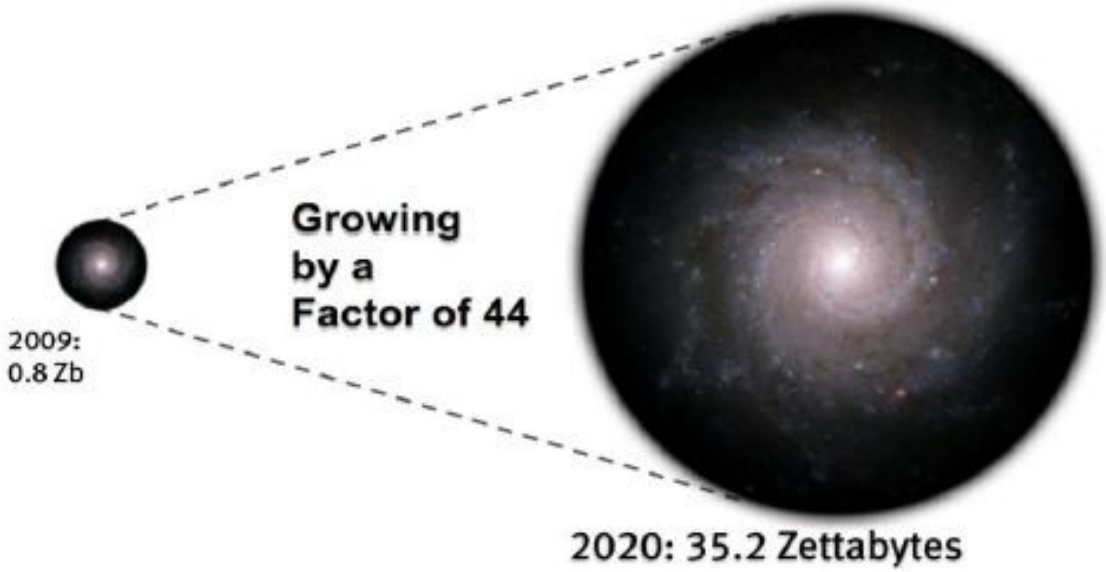
大数据的四个特征



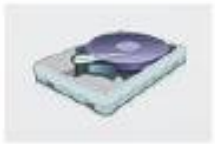
大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

Volume—数量大

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。



数据的度量

TERABYTE	10 的 12 次方	一块 1TB 硬盘		200,000 照片或 mp3 歌曲
PETABYTE	10 的 15 次方	两个数据中心机柜		16 个 Blackblaze pod 存储单元
EXABYTE	10 的 18 次方	2,000 个机柜		占据一个街区的 4 层数据中心
ZETTABYTE	10 的 21 次方	1000 个数据中心		纽约曼哈顿的 1/5 区域
YOTTABYTE	10 的 24 次方	一百万个数据中心		特拉华州和罗德岛州

进入大数据时代

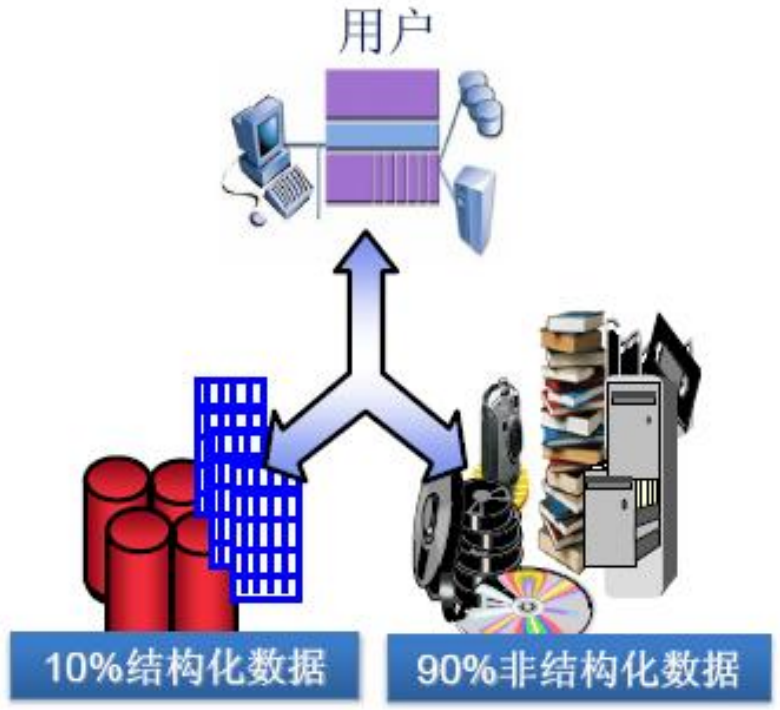
- 2011年，中国互联网行业持有数据总量达到1.9EB（1EB字节相当于10亿GB）
- 2011年，全球被创建和复制的数据总量为1.8ZB（1.8万亿GB）
- 2013年，我们生成这样规模的信息量只需10分钟
- 2015年，全球被创建和复制的数据总量将增长到8.2EB以上
- 2020年，全球电子设备存储的数据将暴增30倍，达到35ZB

Velocity—速度快

- 从数据的生成到消耗，时间窗口非常小，可用于生成决策的时间非常少
- 1秒定律：这一点也是和传统的数据挖掘技术有着本质的不同。

Variety—多样化

- 大数据是由结构化和非结构化数据组成的
  - 10%的结构化数据，存储在数据库中
  - 90%的非结构化数据，它们与人类信息密切相关
- 非结构化数据类型多样
  - 邮件、视频、微博
  - 位置信息、链接信息
  - 手机呼叫、网页点击
  - “长微博”



Value—价值化

- 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，但是具有很高的商业价值
  - 科学研究
  - 企业应用
  - 社会网络

《大数据时代》作者舍恩伯格提出的三个特征

- 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。

大数据的产生

- 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。
- **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，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，这种数据的产生方式是被动的
- **用户原创内容阶段**
  - 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
  - 以博客、微博为代表的新型社交网络的出现和快速发展
  - 以智能手机、平板电脑为代表的新型移动设备的出现
  - 这个阶段数据的产生方式是主动的
- **感知式系统阶段**
  - 感知式系统的广泛使用
  - 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

大数据的应用

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

大数据作用

- **变革价值的力量**
  - 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
- **变革经济的力量**
  - 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
- **变革组织的力量**
  - 大数据将推动网络结构产生无组织的组织力量

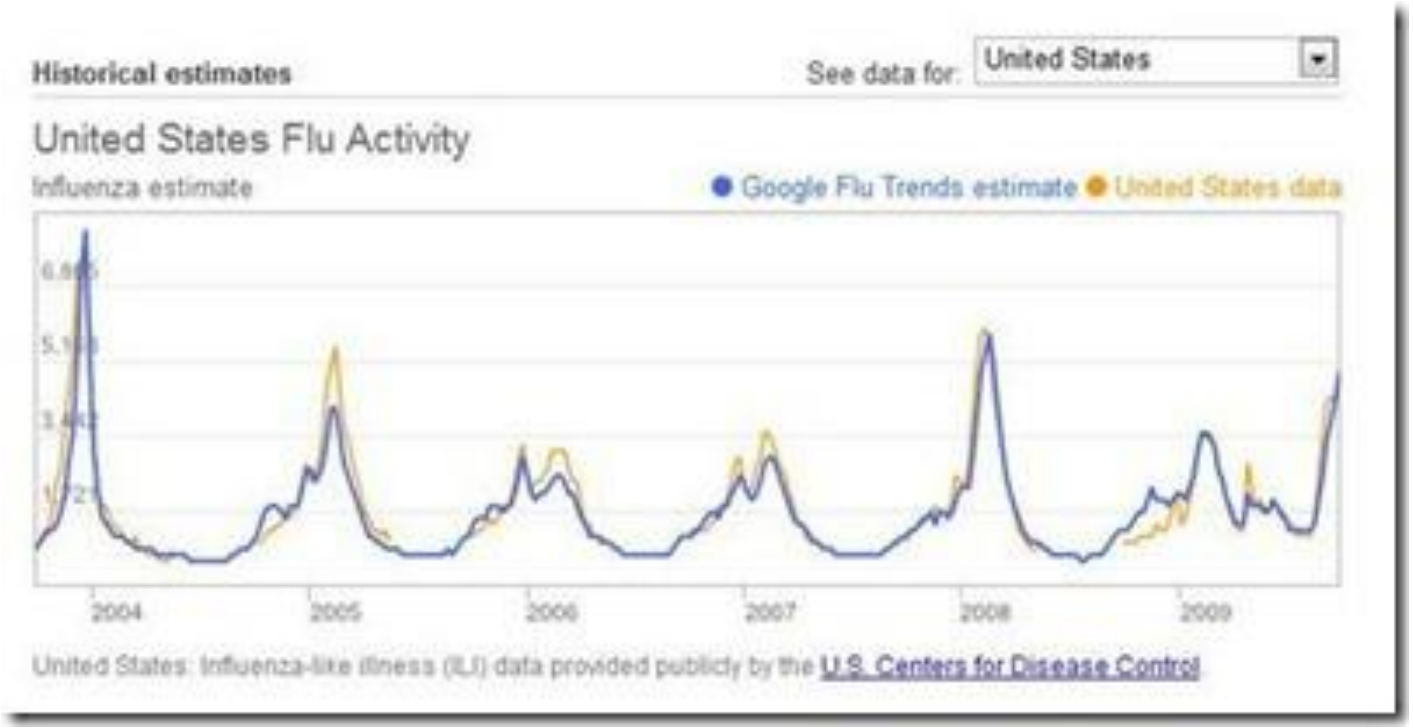
1.4 大数据与大规模数据、海量数据的差别

- 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。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。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。

1.5 典型的大数据应用实例

- 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）



1.6 从数据库到大数据

- 池塘捕鱼（数据库）vs.大海捕鱼（大数据）
- 1、**数据规模：**“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。

1.6 从数据库到大数据

- 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。

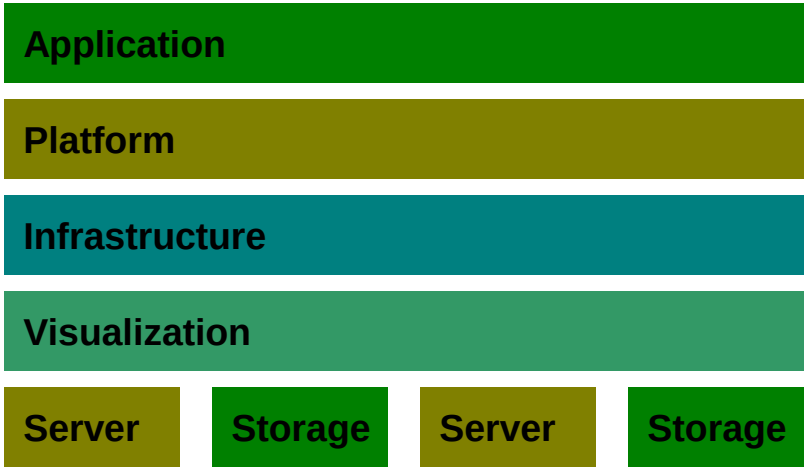
科学研究四种范式

- 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)。

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	Today	Data captured by instruments or generated by simulator; Processed by software;
(eScience)		Information stored in computer; Scientist analyzes database

大数据与云计算

SaaS	从一个集中的系统部署软件，使之在一台本地计算机上(或从云中远程地)运行的一个模型。由于是计量服务，SaaS 允许出租一个应用程序，并计时收费
PaaS	类似于 IaaS，但是它包括操作系统和围绕特定应用的必需的服务
IaaS	将基础设施(计算资源和存储)作为服务出租



SaaS	Software as a Service	Google Apps, Microsoft “Software+Services”
PaaS	Platform as a Service	IBM IT factory, Google App Engine, Force.com
IaaS	Infrastructure as a Service	Amazon EC2, IBM Blue Cloud, Sun Grid
dSaaS	data Storage as a Service	Nirvanix SDN, Amazon S3, Cleversafe dsNet

1.7 大数据与云计算

- **从整体上看，大数据与云计算是相辅相成的**
- **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。

	云计算技术	描述
	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
大数据的关键技术	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

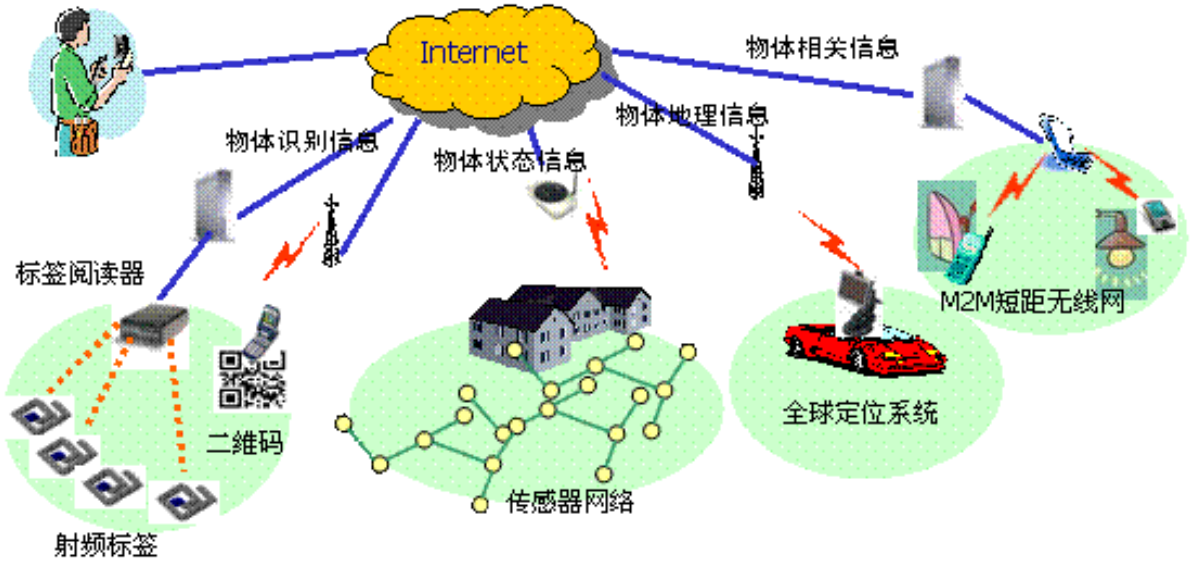
大数据技术与云计算有相同，也有差异

		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

大数据技术与云计算相结合会带来什么

1.8 大数据与物联网

- 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
- 物联网架构可分为三层，包括感知层、网络层和应用层
- 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。

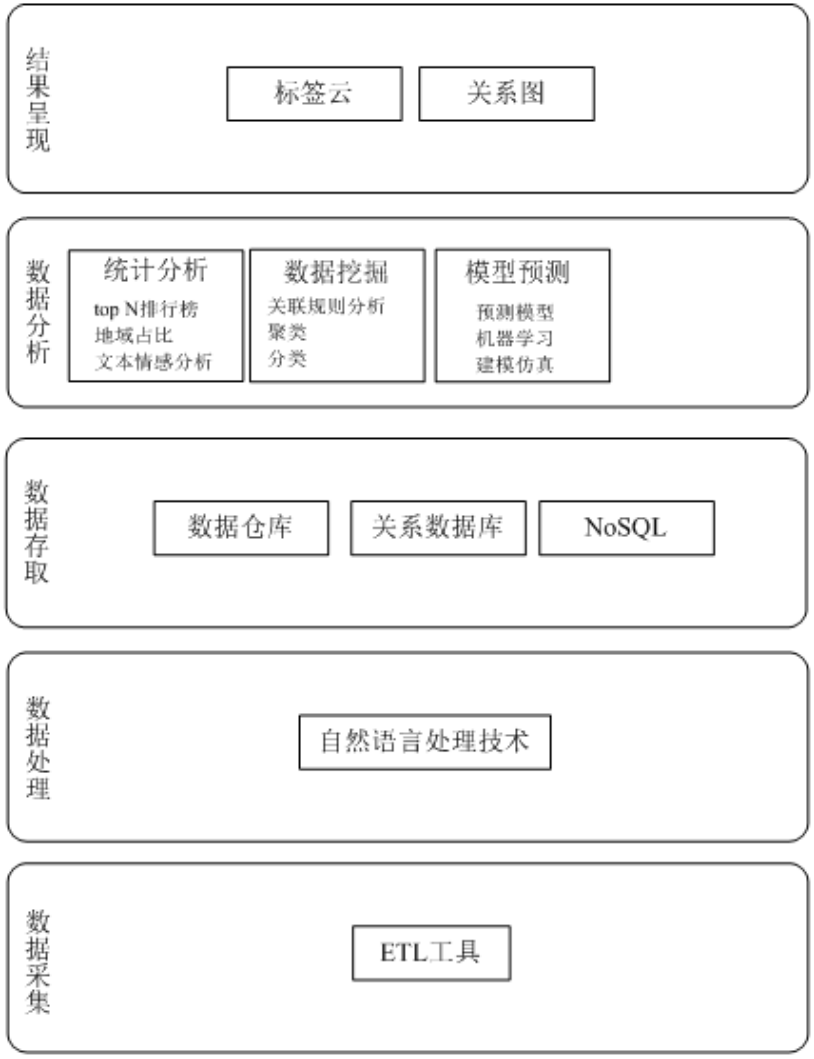


1.9 对大数据的错误认识

- 根据IDC2011年市场研究报告，主要有三个典型的错误说法：
  - 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；
  - 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；
  - 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。

大数据技术

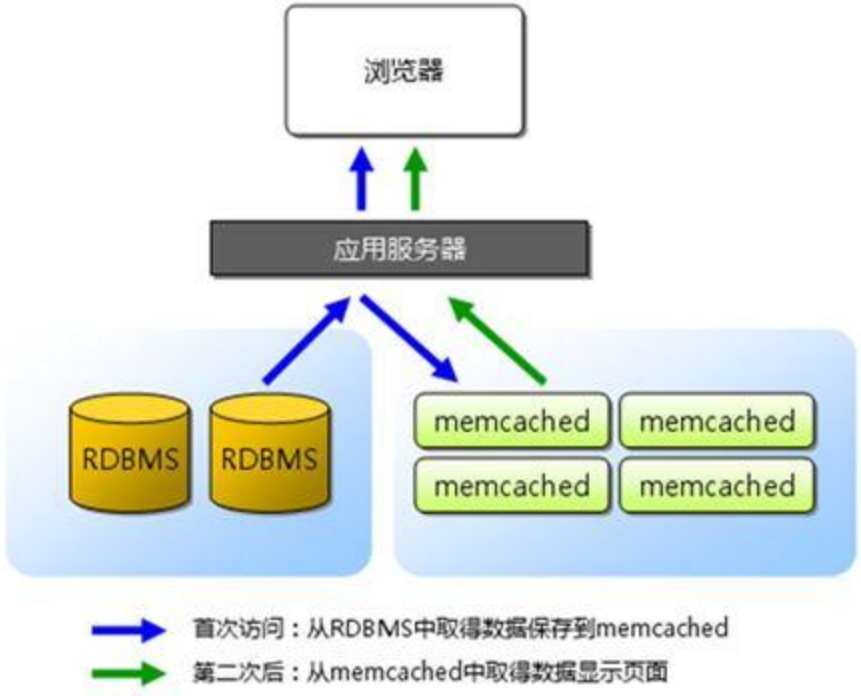

结果呈现
标签云
关系图
数据分析
统计分析
top N排行榜
地域占比
文本情感分析
数据挖掘
关联规则分析
聚类
分类
模型预测
预测模型
机器学习
建模仿真
数据存取
数据仓库
关系数据库
NoSQL
数据处理
自然语言处理技术
数据采集
ETL工具

大数据存储和管理技术

- 大数据时代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。

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。
- 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。

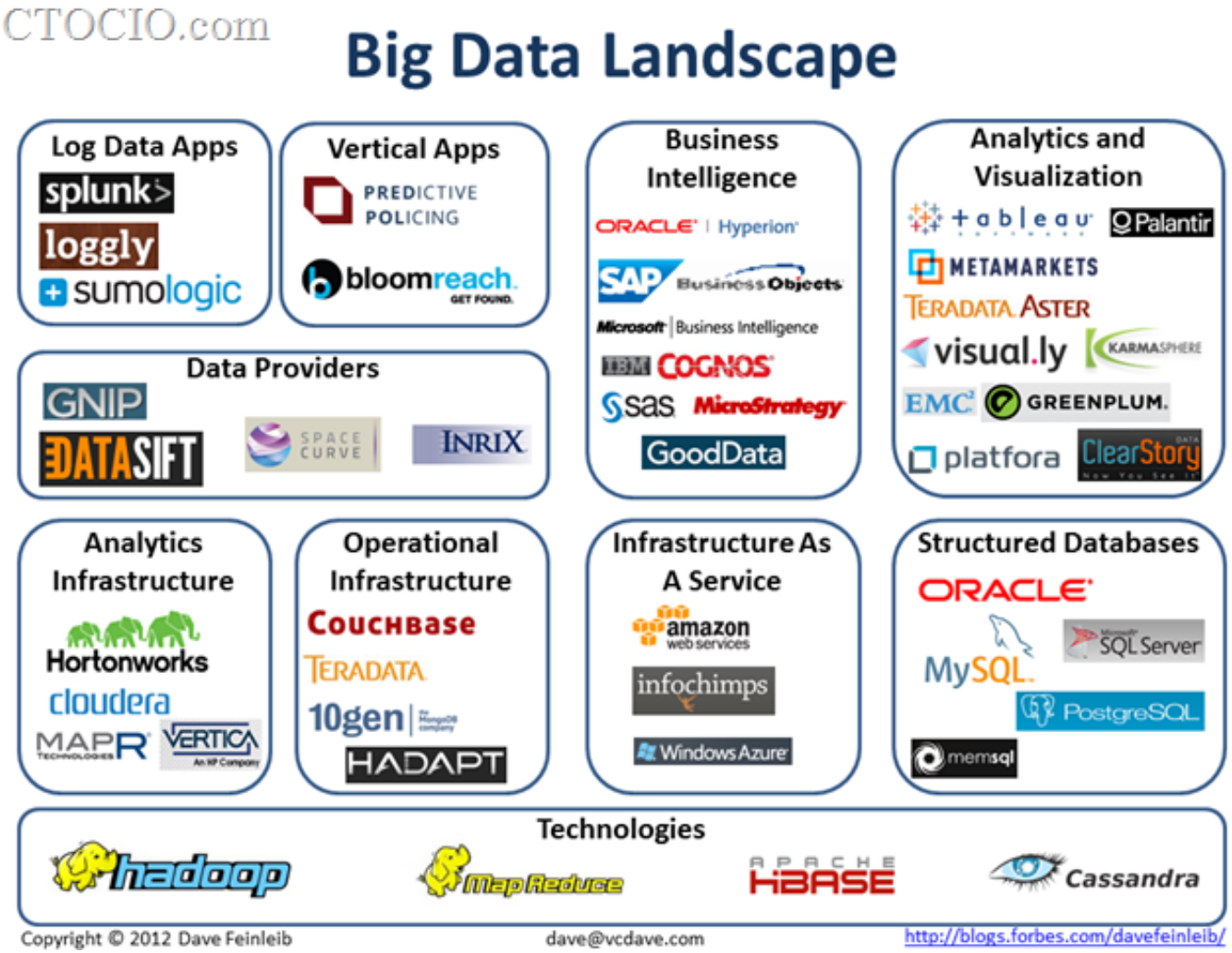
分布式文件系统

- 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。

NoSQL

- 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。

1.12 大数据生态系统


CTOCIO.com
Big Data Landscape
Log Data Apps
splunk>
loggly
sumologic
Vertical Apps
PREDICTIVE POLICING
bloomreach
GET FOUND.
Business Intelligence
ORACLE | Hyperion
SAP BusinessObjects
Microsoft | Business Intelligence
IBM COGNOS
SAS MicroStrategy
GoodData
Analytics and Visualization
tableau
Palantir
METAMARKETS
TERADATA ASTER
visual.ly
KARMASPHERE
EMC² GREENPLUM.
platfora
ClearStory
Data Providers
GNIP
DATASIFT
SPACE CURVE
INRIX
Analytics Infrastructure
Hortonworks
cloudera
MAPR
VERTICA
Operational Infrastructure
COUCHBASE
TERADATA
10gen
HADAPT
Infrastructure As A Service
amazon web services
infochimps
Windows Azure
Structured Databases
ORACLE
MySQL
SQL Server
PostgreSQL
memsql
Technologies
hadoop
MapReduce
APACHE HBASE
Cassandra
Copyright © 2012 Dave Feinleib
dave@vcdave.com
http://blogs.forbes.com/davefeinleib/

主讲教师和助教

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

助教：赖明星

单位：厦门大学计算机科学系数据库实验室2011级硕士研究生（导师：林子雨）
E-mail: mingxinglai@gmail.com
个人主页：http://mingxinglai.com

欢迎访问《大数据技术基础》2013班级网站：http://dblab.xmu.edu.cn/node/423

Department of Computer Science, Xiamen University, Sep, 2013